8-2013

Predictive Validation of the Monitoring Instructional Responsiveness: Reading (MIR:R): Investigation of a Group-Administered, Comprehension-Based Tool for RTI Implementation

Kelli Caldwell Miller
kcaldwe9@utk.edu

To the Graduate Council:

I am submitting herewith a dissertation written by Kelli Caldwell Miller entitled "Predictive Validation of the Monitoring Instructional Responsiveness: Reading (MIR:R): Investigation of a Group-Administered, Comprehension-Based Tool for RTI Implementation." I have examined the final electronic copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in School Psychology.

R. Steve McCallum, Major Professor

We have read this dissertation and recommend its acceptance:

Sherry M. Bell, Christopher H. Skinner, Amy D. Broemmel

Accepted for the Council:
Dixie L. Thompson

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

Predictive Validation of the Monitoring Instructional Responsiveness: Reading (MIR:R): Investigation of a Group-Administered, Comprehension-Based Tool for RTI Implementation

A Dissertation Presented for the

Doctorate of Philosophy

Degree

The University of Tennessee, Knoxville

Kelli Caldwell Miller

August 2013

Acknowledgements

This project would not have been possible without the guidance and support of Dr. R. Steve McCallum and Dr. Sherry Mee Bell. I thank them for the opportunity to work on this project. Additionally, I wish to thank them for their technical and editing suggestions offered during the writing of this dissertation. Their expertise in testing, psychometrics, and teaching has been invaluable. I also wish to thank my committee members, Dr. Christopher Skinner and Dr. Amy Broemmel, for their suggestions, guidance, and discussions about this project.

Additionally, I wish to thank the members of Dr. McCallum and Dr. Bell's research group who participated in this project. Primarily, I thank Dr. Angela Hilton-Prillhart and Dr. Michael Hopkins for the development of the MIR probes. I also wish to thank Elizabeth Hays, Jeremy Coles, Anna Jo Auerbach, and Christy Lyons for their assistance in entering and/or checking probe data.

Steve Duncan, Robyn O'Dell, and Gary Aytes were instrumental in the implementation, data scoring, and data entry of the MIR:R data as well as TCAP data. I also wish to thank David Baldwin for his assistance with database-related information. I appreciate their willingness to share information so that this project may be completed.

Finally, I wish to thank my family for providing support and encouragement throughout the years, as well as during the time required to complete this dissertation. In particular, I wish to thank my husband for providing immeasurable support and understanding.

Abstract

Monitoring Instructional Responsiveness: Reading (MIR:R), a group-administered measure designed to determine at-risk status for reading fluency and comprehension, was administered to 494 third-grade students to determine the relationship between MIR:R static and slope scores and student performance and non-proficiency status on the Tennessee Comprehensive Assessment Program (TCAP) Achievement Test, reading composite. Correlation coefficients defining the relationship between the total MIR:R static score (the Comprehension Rate score) and student performance and non-proficiency status on the TCAP are .60 ($p < .01$) and .52 ($p < .01$), respectively. When the relationships between MIR:R slope and TCAP performance and non-proficiency status were investigated, weaker correlations were obtained--.22 ($p < .01$) and .20 ($p < .01$), respectively. Results from a step-wise multiple regression equation revealed that the MIR:R Comprehension Percentage component score provided moderate predictive validity for TCAP reading composite performance (9.4% age variance accounted for, $p < .01$); the Total Words Read component score was less predictive (1.1 % age additional variance accounted for, $p < .05$). When MIR:R component scores were entered into a logistic regression analysis, these scores predicted TCAP proficiency and non-proficiency status reasonably well; values ranged from 60% to 88% ($p < .01$). Apparently, both the Comprehension Rate total static score and the Comprehension Percentage component score provide solid predictive accuracy ($p < .01$); the slope and Total Words Read component scores are less predictive. Data support the utility of the MIR:R as a promising, progress-monitoring reading screener within a Response to Intervention/problem-solving model.

Table of Contents

CHAPTER I……..1

LITERATURE REVIEW…….1

Relevant Legislation and Implications…….2

Expected Growth on CBM Measures…..……..5

Using CBM Measures as Predictors of Student Achievement…….9

Using Slopes to Illustrate Student Progress……13

Frequency of Administration of Progress-Monitoring Measures……...14

CBM Within Response to Intervention Models…….15

Utility of the Monitoring Instructional Responsiveness: Reading…….23

Statement of the Problem……...25

CHAPTER II…….27

METHOD……..27

Participants……27

Instruments……27

Procedures…….31

Data Cleaning……33

Slope Determination……33

Analyses……35

CHAPTER III……...37

RESULTS……37

Use of TCAP as Criterion Score……...37

Descriptive Statistics of MIR:R and TCAP Scores..……38

Predictive Validity of the MIR:R Overall Static Score…….……...38

Predictive Validity of the MIR:R Overall Slope Score……40

Relative Predictive Power of the MIR:R Component Scores……...41

Relative Predictive Power of the MIR:R Component Scores and Slope…….43

CHAPTER IV……..45

DISCUSSION……..45

CBM Within Response to Intervention Models……...45

Relationships Among the MIR:R Overall Static Score, Slope Score, and TCAP: Zero-Order Correlational Analyses……..47

Relationships Among MIR:R Component Scores, Slope Score, and TCAP: Multivariate Analyses……...50

Accuracy of MIR:R Predictions to TCAP Non- Proficiency Status…….52

Summary……...53

Limitations and Future Research……..54

REFERENCES…….57

APPENDIX A……...67

VITA…….77

List of Tables

Table 1 Alternate Form Reliability – Monitoring Instructional Responsiveness: Reading (MIR:R)……68

Table 2 Descriptive Statistics and t Tests for Monitoring Instructional Responsiveness: Reading (MIR:R) Slopes……69

Table 3 Descriptive Statistics for Monitoring Instructional Responsiveness: Reading (MIR:R), Sixth Probe Administration, and Tennessee Comprehensive Assessment Program (TCAP), Achievement Test, Reading Composite Scores……70

Table 4 Correlation Coefficients Among Monitoring Instructional Responsiveness: Reading (MIR:R) Comprehension Percentage, Total Words Read, Comprehension Rate Static, Comprehension Rate Slope, and Tennessee Comprehensive Assessment Program (TCAP), Achievement Test, Reading Composite Scores……71

Table 5 Sensitivity and Specificity Information for Monitoring Instructional Responsiveness: Reading (MIR:R) Comprehension Rate Static Score to Predict Tennessee Comprehensive Assessment Program (TCAP), Achievement Test, Reading Composite Non-Proficiency Status……...72

Table 6 Sensitivity and Specificity Information for Monitoring Instructional Responsiveness: Reading (MIR:R) Comprehension Rate Slope to Predict Tennessee Comprehensive Assessment Program (TCAP), Achievement Test, Reading Composite Non-Proficiency Status……73

Table 7 Step-wise Multiple Regression Analysis Predicting Tennessee Comprehensive Assessment Program (TCAP), Achievement Test, Reading Composite, with Monitoring Instructional Responsiveness: Reading (MIR:R) Comprehension Percentage and Total Words Read Component Scores……..74

Table 8 Sensitivity and Specificity Information for Monitoring Instructional Responsiveness: Reading (MIR:R) Comprehension Percentage and Total Words Read Component Scores to Predict Tennessee Comprehensive Assessment Program (TCAP), Achievement Test, Reading Composite Non- Proficiency Status……...75

Table 9 Step-wise Multiple Regression Analysis Predicting Tennessee Comprehensive Assessment Program (TCAP), Achievement Test, Reading Composite, with Monitoring Instructional Responsiveness: Reading (MIR:R) Comprehension Percentage, Total Words Read, and Comprehension Rate Slope…….76

CHAPTER I

Increasingly, educators are held accountable for students' academic progress. In an effort to improve assessment and monitoring of progress, educators are relying more on formative evaluation, particularly within a Response-to-Intervention (RTI) framework. Research on the use of formative evaluations within the RTI paradigm indicates that frequent measurement leads to improved student outcomes and teacher planning (Fuchs, Deno, & Mirkin, 1984; Fuchs, Fuchs, Hamlett, & Stecker, 1991; Fuchs, Fuchs, Hamlett, Walz, & Germann, 1993; Jones & Krouse, 1988). Most of the existing reading instruments require individual assessment and measure only oral reading. One formative measurement system developed for the assessment of K-5 reading is Monitoring Instructional Responsiveness: Reading (MIR:R; Bell, Hilton-Prillhart, McCallum, & Hopkins, 2012). However, data establishing the psychometric properties of MIR:R are needed, particularly, predictive validity. Consequently, this study was designed to determine the extent to which various MIR:R scores (i.e., Comprehension Percentage, Total Words Read, Comprehension Rate score and slope) predict student performance on a large-scale, end-of-the-year measure, the Tennessee Comprehensive Assessment Program (TCAP) Achievement Test (Tennessee Department of Education, http://tn.gov/education/assessment/achievement.shtml).

LITERATURE REVIEW

Consistent with the purposes of the study, this review of the literature includes a description of the assessment context created by recent educational legislation designed to identify academically at-risk students and the impact of this legislation on assessment practices within school systems. This review also includes descriptions of various strategies used to predict student achievement, particularly Curriculum-Based Measurement (CBM) tools, as well as the growth one might expect on these measures. Next, this review includes a description of

how academic growth may be operationalized by CBM slopes and the literature addressing the optimal frequency of these measures, i.e., how many and how often they should be administered in order to obtain an accurate representation of students' levels of performance. Finally, the review includes a description of some current measures used within the Response to Intervention (RTI) framework, particularly the unique MIR:R (Bell et al., 2012). Many currently-available CBM measures used within the RTI framework operationalize reading fluency as the number of words read correctly within a 1-minute time period; reading comprehension is typically operationalized as student responses to timed reading passages (e.g., selecting words that fit into interspersed blanks, answering content-based questions, or oral retelling of what was read).

Relevant Legislation and Implications

The Individuals with Disabilities Education Improvement Act (IDEIA; Congress, 2004) and No Child Left Behind (NCLB; Congress, 2001) emphasize improving students' academic achievement, particularly the achievement of students who are considered "at risk." The passage of these two educational laws has resulted in schools and state departments of education placing more importance on student outcomes and related indices of accountability. Following these emphases on student outcomes and accountability, the NCLB legislation has advocated for students who appeared to "fall through the cracks." Specifically, NCLB requires that all children make adequate yearly progress toward achieving state-mandated standards. Additionally, NCLB requires that school systems "disaggregate […] data for various groups," in an effort to ensure that all groups of students are making progress (Bell & McCallum, 2008, p. 10). In doing so, NCLB requires that each school system provide evidence of the effectiveness of its educational efforts (Shapiro, Keller, Lutz, Santoro, & Hintze, 2006; Shinn, Shinn, Hamilton, & Clarke, 2002; Thurlow & Thompson, 1999).

Evidence of efforts to improve student progress toward state-mandated standards may be obtained from a variety of sources, including both summative and formative assessments. These assessments may be comprised of varying formats (e.g., end-of-chapter quiz grades, standardized assessments, weekly spelling tests). Typically, summative tests are those given at the end of a unit or at the end of the school year to determine whether a student has mastered content; on the other hand, formative tests are those given throughout a unit or school year to determine students' progress toward state-mandated standards. Ultimately, the results of these assessments may be used to determine whether a student adequately masters basic skills and content knowledge. Additionally, the results of these assessments may be used to predict student performance on end-of-the-year achievement tests.

One strategy that uses summative assessment procedures to predict end-of-year progress requires administration and interpretation of standardized, norm-referenced reading tests. Such tests include: the Test of Silent Word Reading Fluency (TOSWRF; Mather, Hammill, Allen, & Roberts, 2004), the Test of Silent Contextual Reading Fluency (TOSCRF; Hammill, Wiederholt, & Allen, 2006), the Test of Word Reading Efficiency (TOWRE; Torgensen, Wagner, & Rashotte, 1999), the Gray Oral Reading Tests-Diagnostic (GORT-D; Wiederholt & Bryant, 2001), the Nelson-Denny Reading Test (Brown, Fishco, & Hanna, 1993), the Woodcock-Johnson – III Tests of Achievement, Third Edition (WJ-III; Woodcock, McGrew, & Mather, 2001), the Kaufman Test of Educational Achievement, Second Edition (KTEA-II; Kaufman & Kaufman, 2004), AIMSweb© (Shinn & Shinn, 2002), and the Dynamic Indicators of Basic Early Literacy Skills (DIBELS; Good & Kaminski, 2002).

Administration formats for these assessments vary. Additionally, the aspects of reading that are measured by each assessment vary. Therefore, these assessments may provide different

estimates of prediction, based upon the unique aspects of reading assessed as well as the administration format. Consequently, determining the "optimal" measure is difficult and depends upon a variety of factors (e.g., administration time available, match between content of the predictor and criterion measures). For instance, only a few of the aforementioned assessments (e.g., TOSWRF, TOSCRF, and Nelson-Denny) allow for a group administration format; the rest are conducted using an individual administration format, which can be very inefficient and time consuming. In addition, they do not provide multiple alternate forms. Consequently, these assessments are typically administered only to children who are referred for a comprehensive, psychoeducational assessment and, in particular, to those who may be considered for eligibility for special education, rather than to monitor ongoing progress.

For those school systems that emphasize the value of predicting end-of-year progress and determining the extent to which *all* students are likely to achieve state-mandated standards, efficiency of test administration and utility of assessment results become paramount issues. If a particular reading assessment is inefficient, educators may be less likely to use that particular reading instrument. In addition, if the information is limited (e.g., it provides information only about a student's reading rate, rather than comprehension or both rate and comprehension), educators may opt for a different assessment, depending on the purpose of the assessment. Potential benefits of the test (e.g., types and amount of information obtained) must be weighed against the time required for administration to all students, rather than a select few. Therefore, assessments should be efficient, yield high-quality, desired information, and be fairly easy to administer and interpret.

The passage of the IDEIA legislation influenced adoption of ongoing monitoring of progress within a formative evaluation format because of its emphasis on use of an RTI model to

gauge student mastery, and ultimately, eligibility for special education services. Taken together, the legislative efforts of both IDEIA and NCLB have influenced schools' and teachers' assessment practices and have influenced both formative and summative assessment, resulting in changes in methods of evaluating student progress toward state-mandated academic standards.

Formative assessment evolved from formative evaluation (Dorn, 2010; Scriven, 1967). As Scriven (1967) conceptualized it, the intent of formative evaluation is to gather information in an attempt to evaluate the effectiveness of a curriculum as well as to guide instructional changes. Over time, formative assessment has been expanded and adapted to the context in which it is currently used (i.e., as Curriculum-Based Measures within a Response-to-Intervention paradigm). Optimally, within classroom settings, formative assessment may resemble a feedback loop. For instance, instructional changes are made as a result of student performance. This process is repeated (i.e., following some time implementing the new instructional changes, student performance is again assessed and instruction may be further modified to address the needs of lower-performing students). Ultimately, formative assessment may be used to alter subsequent educational decisions (Wiliam, 2006). In the current context of the RTI paradigm, formative assessment is used to determine whether empirically-validated interventions are improving student performance on Curriculum-Based Measurement (CBM) tests. CBM tests are administered as formative, basic-skill assessments designed to monitor students' progress (Shapiro, 2004).

Expected Growth on CBM Measures

Although CBM has been researched for many years, the use of CBM measures to assess universal student progress within the RTI paradigm is a relatively new phenomenon. Consequently, there is considerable interest in determining the best indicators of student growth

using CBM measures. In addition, because educators are interested in improving measurement efficiency, they are increasingly focusing on determining the relationship between growth in student performance and frequent monitoring of student progress. Documenting how quickly students can improve their basic skills, given exposure to certain empirically-validated interventions, is critical.

In an attempt to discern how much growth is possible over time, Fuchs et al. (1993) used students' scores on CBM reading fluency and reading comprehension measures over a two-year period. In order to assess reading fluency in year one, the authors constructed grade-level reading passages and had students read aloud from them under certain time constraints to determine the number of words read correctly per minute. In order to assess reading comprehension in year two, students engaged in computer-administered CBM maze procedures, which involved reading a passage with words missing. Students were required to select an appropriate word to fit into blank spaces in the passages.

Fuchs et al. (1993) found that "the effect of grade level on CBM reading slopes differed for the two types of measures," and the authors concluded that their findings followed the developmental trajectory of reading (Fuchs et al., 1993, p. 35). In the lower grades, students make the most dramatic gains in the process of acquiring basic reading skills; in contrast, students in higher grades have already acquired many basic reading skills, so their growth is likely to be less steep than those students in the lower grades (Fuchs et al., 1993). Therefore, students in the lower grades are more likely to have greater rates of learning, and show more growth over a shorter amount of time, than students in the higher grades. Additionally, Fuchs et al. (1993) concluded that this difference in growth rates as a function of grade level may be mediated by which type of CBM measure is used to assess student growth. For instance, the

authors concluded that the oral reading of grade-level passages depends on the "component skills proposed by developmental reading theorists: decoding and fluency" (Fuchs et al., 1993, p. 36).

Fuchs et al. (1993) hypothesized that greater growth on the CBM reading fluency tests would be found for students in lower grades, whose reading skills are still developing, than for students in higher grades, whose reading skills are relatively well-established. Fuchs et al. (1993) did, in fact, find this to be true. Alternatively, based on developmental reading theory, the authors hypothesized that the CBM maze tests would require a more comprehensive set of component skills, rather than just decoding and fluency (Fuchs et al., 1993). Therefore, the rates of progress measured by CBM maze tests should be more consistent among the grades, unlike the rates of progress measured by CBM reading fluency tests. Both hypotheses were supported: greater growth rates on the CBM reading fluency test for younger students were obtained, as were more consistent rates of progress for each grade level using the CBM maze test, as opposed to the CBM reading fluency test.

Once all slopes of student progress were calculated, the authors analyzed the data to determine appropriate growth per week, defined as word acquisition, according to each grade level assessed (Fuchs et al., 1993). These growth expectations reflect the developmental nature of reading, with more growth expected in the lower grades and less growth expected in the higher grades. Fuchs et al. determined realistic and ambitious goals for weekly growth, respectively, as:

> "1.5 and 2.0 words per week at Grade 2; 1.0 and 1.5 words per week at Grade 3; .85 and 1.1 words per week at Grade 4; .5 and .8 words per week at Grade 5; and .3 and .65 words per week at Grade 6" (p. 35).

More recently, Jenkins and Terjeson (2011) assessed student growth on CBM reading passages. Similar to Fuchs et al. (1993), Jenkins and Terjeson (2011) assessed the number of words read correctly (WRC) under certain time constraints on grade-level passages to determine WRC per minute. Additionally, this particular sample included students at varying grade levels who were receiving special education services. Jenkins and Terjeson (2011) found an increase in the mean number of WRC per minute at each measurement point. During the eight weeks of the study, the authors calculated "a mean growth slope between 1.48 and 1.67 WRC per week" (Jenkins & Terjeson, 2011, p. 33). The authors concluded that their results supported those obtained by previous researchers (Deno, Fuchs, Marston, & Shin, 2001), suggesting that a growth of approximately 1.5 words per week may be considered realistic for students at varying grade levels receiving evidence-based interventions (Jenkins & Terjeson, 2011). Although Fuchs et al. (1993) determined different goals for each grade, Jenkins and Terjeson (2011) determined that one consistent goal across all grades may be appropriate. This difference may be due to the influence of evidence-based interventions that the students in Jenkins' and Terjeson's (2011) study received, in contrast to more traditional instruction, which students in the Fuchs et al. (1993) study received.

It appears as though a steady rate of growth can be expected for students receiving adequate, evidence-based instruction. However, the developmental trajectory of acquiring basic reading skills likely mediates student growth over time. Educators may be able to expect a greater rate of growth for students in lower grades, while their basic reading skills are still developing, on CBM reading fluency measures. Students in higher grades may be less likely to evidence such increased growth rates on CBM reading fluency measures since their basic reading skills are relatively well established. However, these particular students may show

increased growth on comprehension measures, such as CBM maze measures. In order to obtain more consistent rates of progress among students in varying grade levels, Fuchs et al. (1993) suggest the use of CBM maze procedures, which require a more comprehensive set of basic reading skills, rather than just decoding and fluency. Of course, the maze procedure is just one type of comprehension assessment strategy, and other strategies might be more (or less) effective.

Using CBM Measures as Predictors of Student Achievement

In addition to informing instruction, another important goal for educators is to predict student performance on end-of-the-year achievement tests using a range of variables (e.g., student grades, informal and formal assessments, standardized and nonstandardized measures). However, using these options may limit researchers and/or teachers to specific points in time in which these data can be collected and may limit the use of a more immediately-available score and/or multiple scores. With the recent implementation of RTI models, school personnel are accumulating multiple measures of progress (probes) over the course of an academic year. However, research regarding the utility of these more immediately-available student scores and, particularly, their power to predict student performance on end-of-the-year achievement tests, is limited (Crawford, Tindal, & Stieber, 2001; Helwig & Tindal, 1999; Nolet & McLaughlin, 1997, Shapiro et al., 2006; Silberglitt & Hintze, 2005).

Researchers who have conducted studies investigating the predictive utility of CBM tests have found positive results (Crawford et al., 2001; Shapiro et al., 2006; Silberglitt & Hintze, 2005). For example, Crawford et al. (2001) used reading rate performances on CBM reading assessments to predict student performance on statewide achievement tests. Using passages constructed from Houghton Mifflin Basal Reading Series, Crawford et al. (2001) had students

read three passages for 1 minute each. Following completion of the third passage, the authors averaged the number of words read per minute (reading rate) across the three passages, obtaining a reading rate score from students in second grade during the first year of the study, and then obtaining a reading rate score from the same students, now in third grade, during the second year of the study. In addition to comparing the stability of students' scores between year one and year two of the study, Crawford et al. (2001) correlated students' reading rate scores to their scores on the Oregon Department of Education's end-of-the-year reading achievement test. When using the averaged score from the CBM reading assessments as the predictor variable, over 80% of the students reading at the 50th percentile and above were able to pass the statewide reading achievement test when both second-grade and third-grade data were analyzed (Crawford et al., 2001). When analyzing the across-years data, the authors found that 100% of students who obtained a reading rate of 72 words per minute or more in second grade were able to pass the statewide reading achievement test in third grade (Crawford et al., 2001).

Similarly, Shapiro et al. (2006) investigated the utility of CBM reading assessments to predict students' performance on end-of-the-year state achievement tests. Shapiro et al. (2006) also used reading rate performances on CBM assessments to predict student performance on the Pennsylvania System of School Assessment (PSSA) reading achievement test. Using passages from the AIMSweb© assessment system and another independently-created CBM assessment system, the authors required students to read one passage for 1 minute, obtaining a reading rate for each student (Shapiro et al., 2006). Shapiro et al. (2006) obtained moderate correlation coefficients between both CBM assessment systems and the PSSA reading achievement test (range of .62-.69, with the exception of one lower correlation of .25).

Likewise, Silbertglitt and Hintze (2005) examined correlations between student scores on CBM reading assessments and the Minnesota Comprehensive Assessment (MCA) reading achievement test, but also incorporated cut scores for student performance on CBM assessments into their analyses. They contend that cut scores on CBM assessments allow educators to identify those students who are performing at proficient levels for their currently-developing reading skills (e.g., phonetic decoding, fluency, and comprehension). Silberglitt and Hintze (2005) speculated that those students whose scores fell above the cut score would be able to pass the state reading achievement test, whereas those students whose scores fell below the cut score would not be able to pass the state reading achievement test. Students were assessed on three CBM passages three times during the year: fall, winter, and spring; the final score obtained during each assessment period was the median score of the three reading passages. The authors used data collected during the spring administration of the CBM reading assessment to predict students' scores on the state reading achievement test. To determine cut scores, Silberglitt and Hintze (2005) used a combination of logistic regression and ROC curve analysis. Confirming their hypothesis, the authors found that a high percentage of those students whose scores fell above the pre-determined cut score, defined as the minimal score that students may obtain in order to be considered proficient in reading on the CBM assessment (e.g., the minimal number of words read correctly per minute considered to be proficient) also passed the end-of-the-year state achievement test (Silberglitt & Hintze, 2005).

Conversely, when using slope as a predictor, other researchers have found that CBM offers little predictive validity for future performance on state achievement tests or other future reading performance (Schatschneider, Wagner, & Crawford, 2008; Stage & Jacobsen, 2001; Yeo, Fearrington, & Christ, 2012). Yeo et al. (2012) investigated the predictive validity of

different types of CBM reading measures, AIMSweb© oral reading fluency and maze reading, specifically, to predict end-of-the-year scores on the Tennessee Comprehensive Assessment Program (TCAP) Achievement Test, reading composite. Using a bivariate latent growth modeling technique, the authors found that the slope estimates provided by student performance on the oral reading fluency measures were not significantly correlated with the estimates provided by student performance on the maze reading measures. Additionally, Yeo et al. (2012) concluded that CBM growth estimates did not contribute to the predictions of student performance on the reading composite on the end-of-the-year state achievement test, the TCAP. The authors indicate that the lack of predictive utility of the CBM scores may be due, in part, to the unstable nature of a growth rate derived from multiple data points, or the instability of slope. Similarly, Schatschneider et al. (2008), as well as Stage and Jacobsen (2001), found that CBM slopes offered little predictive power for students' future reading performance.

Perhaps CBM predictive utility is limited as a function of the particular scores used (or the predictor used). Silberglitt and Hintze (2007) used hierarchical linear modeling to establish and compare student growth rates based upon initial level of performance. The authors investigated student performance over time, rather than use the students' initial growth rates to predict to a certain criterion. In this study, students were assessed using the standardized procedures of CBM (Shinn, 1989) and data were analyzed using all three benchmarks (e.g., fall, winter, and spring). The authors grouped students, by grade level, "based on the normative ranking of their fall […reading] CBM score" (Silberglitt & Hintze, 2007, p. 74). Student growth rates were compared across all ten deciles.

The authors found that growth rates, defined as slopes, varied significantly among students and decile groups (Silberglitt & Hintze, 2007). Specifically, results show that the growth

rates for those students in the lowest and highest deciles were much lower than the growth rates for students in the middle deciles. This is not surprising, given that students in the lowest deciles are struggling to develop reading skills, whereas students in the highest deciles may be already performing at proficient levels and "top out." Both scenarios reduce variability and the range of possible scores. Reduced range of scores limits the magnitude of correlation coefficients. According to Silberglitt and Hintze (2007), it may be difficult to obtain an accurate representation of skill growth for students who are in the lowest- and highest-performing groups, when compared to their peers.

In summary, because of the current legislative emphasis on accountability for all students, educators are increasingly interested in using currently-available measures of student progress to predict students' end-of-the-year, state achievement test performance. In order to glean information about each student's current performance level, educators need time-efficient, easy-to-administer assessments that will provide the types of information in which they are interested. In the case of reading, educators need information about both reading fluency and reading comprehension. However, few of the currently-available assessment instruments measure both reading fluency and reading comprehension in a time-efficient manner. Many require multiple subtests to be administered to obtain information regarding both skills. Therefore, MIR:R has been developed as an efficient, easy-to-administer, Curriculum-Based Measurement tool.

Using Slopes to Illustrate Student Progress

Because of the current legislative emphasis on frequent probe administration to monitor students' progress, within the RTI paradigm, CBM measures allow educators to obtain multiple scores within a short period of time. Once the data are obtained how might they be most

efficiently used? One way to incorporate all currently-available assessment data points into a cohesive measurement unit is to calculate a slope for each student.

Many studies that have investigated the utility of CBM measures have illustrated student performance within a slope format, rather than individual scores obtained during each of the various administration sessions (Deno et al., 2001; Good, Deno, & Fuchs, 1995; Good & Shinn, 1990; Kim, 1993; Shin, Deno, & Espin, 2000; Shin, Deno, McConnell, & Espin, 2000). These studies provide support for conceptualizing growth assessed by CBM measures as a linear function (Deno et al., 2001). For instance, students may have their progress assessed every week. In this case, one could measure a weekly rate of progress over an entire grading period.

Additionally, slopes reflect the dynamic nature of skill acquisition. Slopes are sensitive to minor changes in skill acquisition over short periods of time. Therefore, slopes are appropriate illustrations of student progress when considering the developmental trajectory of basic skills. Because slopes allow observation of similar data along a linear continuum, the rate of change, or growth over time, can be determined and then compared to peers who may be receiving the same intervention. Slopes may also be used to predict future performance on a variety of assessment techniques (e.g., end-of-unit-tests, progress-monitoring measures, and end-of-the-year state achievement tests) because they provide a concrete representation of a student's rate of growth over time.

Frequency of Administration of Progress-Monitoring Measures

Because growth on progress-monitoring measures may be small within short time intervals, educators must carefully consider the amount of time spent on administration of progress-monitoring measures and weigh it against the amount of growth the student is likely to make within a specified time interval. If the growth rate is expected to be an approximate

increase of 1.5 WRC every week (Jenkins & Terjeson, 2011), then is it necessary to administer progress-monitoring probes weekly? Additionally, if instructional changes are to be informed by student performance on progress-monitoring measures and administration of these measures is to be conducted weekly, can teachers readily determine which instructional changes are improving or failing to improve student performance?

Jenkins and Terjeson (2011) investigated the frequency of administering progress-monitoring probes to students. The time intervals between each administration varied: every 2, 4, or 8 weeks. The authors found that the slopes of student performance when the progress-monitoring probes were administered using the every-two-week schedule were highly correlated to the slopes of student performance when the progress-monitoring probes were administered on the every-eight-week schedule. Not only were the slopes from varying time intervals highly correlated, but the slopes were also similar in magnitude (i.e., means between 1.48 and 1.67 WRC per week) across all schedules (i.e., every 2, 4, or 8 weeks), which is consistent with previous research on the frequency of progress monitoring (Jenkins, Graff, & Miglioretti, 2009).

Although frequent progress monitoring is beneficial to determine whether students are responding to empirically-validated instruction, the quality of information gleaned from these measures must be weighed against the amount of time required to complete the administration of these measures. If a teacher can obtain similar data as a function of assessments taken over longer time intervals, then the time gained from fewer administrations may be used to provide more instructional services to students.

CBM Within Response to Intervention Models

Currently, response to evidence-based interventions is recognized as one criterion to define academic progress and to consider in diagnosing specific learning disabilities. Rate is

defined as progress on CBM-type measures, which are amenable to use within an RTI problem-solving paradigm. Initially, CBM tests were designed to measure progress within a specified curriculum; however, current versions of CBM tests are not likely to be based in the curriculum (Bell & McCallum, 2008; Fuchs & Deno, 1994). Fuchs and Deno (1994) sought to determine whether testing material must be drawn from students' instructional curricula by reviewing the available literature for the extent to which using instructional curricula exerts a controlling effect on instructional decisions as well as the advantages and disadvantages associated with using students' instructional curricula. Fuchs and Deno (1994) found that passages randomly selected from the curriculum were no more effective for defining student gain than other grade-specific passages. Nor were curriculum-based passages more psychometrically robust (Fuchs & Deno, 1994). Thus, Fuchs and Deno (1994) concluded that material that comprises curriculum-based measures does not have to come from students' instructional curricula. Results of this study produced a shift in the field, away from testing material comprised of curriculum-specific content, and toward more generic, grade-level content.

Consequently, many current CBM tests are described as hybrids, combining features of formal and informal assessments (Bell & McCallum, 2008). These measures are sometimes chosen over standardized tests because standardized tests are typically more time-consuming, administered on an individual basis, and are not as sensitive to small changes in student achievement. Moreover, many CBM reading measures have multiple forms, which allow for frequent monitoring of student progress over long periods of time. Additionally, research indicates that CBM reading measures are sensitive to growth and correlate well with standardized measures of reading achievement (Deno et al., 2009; Marston, 1989).

Alternatively, many CBM measures are criticized for not providing educators with relevant information concerning a student's ability to decode and comprehend authentic, connected text (Bell & McCallum, 2008; Brunsman & Shanahan, 2006; McGill-Franzen, Dennis, Payne, & Solic, 2006). For instance, requesting that a student say the names of letters embedded in a string of randomly-generated letters does not give the evaluator information about what letter sounds that student knows. That is, most current CBM measures gauge progress but do not give information about mastery of specific skills. Therefore, the utility of CBM measures to provide relevant information that can be used within a classroom setting must also be considered when determining the type of CBM measure to be used in a problem-solving context.

Some CBM measures are also criticized for a lack of predictive accuracy (Johnson, Jenkins, Petscher, & Catts, 2009). Passages that comprise CBM reading measures are typically brief and vary in difficulty level (Poncy, Skinner, & Axtell, 2005), which may lead to decreased ability to accurately predict to criterion measures as well as limited reliability. Finally, some experts (e.g., Poncy et al., 2005) suggest the administration of multiple probes to determine a student's current performance, rather than relying on score(s) obtained from a single probe, given the limited reliability of the single, brief measures. Johnson et al. (2009) suggest the calculation of sensitivity (true positives) and specificity (true negatives) indices to determine how accurately a CBM measure predicts future performance that is defined categorically.

Within an RTI model, CBM measures are used to identify those students who are struggling, often conceptualized as those performing within the lowest 10% of the students within a specific grade level. This occurs after every student in the grade has been assessed, using varying CBM measures, depending on what types of measures the school or school system has adopted. These various measures may require an individual- or group-administration format.

Depending on the administration format and the number of measures to be administered, the time required to administer the measures to all students may be quite lengthy. Additionally, if the assessment system used by the school requires multiple measures to obtain an adequate representation of a student's reading ability, this can prove to be quite costly.

Once the lowest 10% have been identified, these students may receive empirically-validated interventions for a certain amount of time per week in an effort to improve their basic-skill deficiencies within a tiered methodology. During this intervention period, students' progress toward acquisition of basic skills is monitored via the CBM measures. If a student progresses at a certain rate, sometimes defined as the rate of progress of the student at the 25th percentile, he/she may no longer receive the special intervention and return to general education services (e.g., Tier 1). However, if the student does not progress at a certain rate and continues to remain in the lowest echelon of his/her peers, then more intense interventions are provided (e.g., Tier 3). It is only after a student fails to respond to increasing levels of empirically-validated interventions that special education eligibility may be considered.

Because of the emphasis on monitoring students' progress within the RTI paradigm, CBM measures must be efficient and cost effective. However, not all of the currently-available measures are both efficient and cost effective, while still providing all of the desired information regarding a student's progress. Of the CBM measures currently being used in school settings, DIBELS (Good & Kaminski, 2002) and AIMSweb© (Shinn & Shinn, 2002) are the most popular. Both of these CBM assessment systems provide scores reflecting a student's reading ability. DIBELS and AIMSweb© both include oral reading fluency measures as the main measure of a student's reading ability. Administration of the oral reading fluency measure for both assessment systems requires a student to read three 1-minute passages. The examiner then

takes the median score for the number of words read correctly in 1 minute and takes the median score for the number of errors made in 1 minute. Thus, the examiner can obtain information concerning how many words a student can read correctly within a 1-minute time period, as well as information concerning how many errors a student may make within a 1-minute time period. According to Shinn (2002), the basic decision-making metric for using CBM oral reading fluency measures within a decision-making/problem-solving model requires determining the number of words read correctly in a 1-minute period. Shinn (2002) also purports that best practices for using a CBM maze measure within a decision-making/problem-solving model include using the number of correct word choices in a 5-minute period.

Currently-available instruments may measure aspects of reading fluency or reading comprehension, or, possibly, both. Researchers may differ on the level of importance placed upon reading fluency and reading comprehension. McKenna and Stahl (2003) indicate that fluency involves accuracy, automaticity, and appropriate inflection/prosody. Fluency may also be conceptualized as one's ability to read a passage with appropriate rhythm and phrasing and is evidenced by good readers when they engage in oral reading. In terms of assessment, fluency is typically conceived of as the number of words read correctly in 1 minute. In this conceptualization of reading fluency, some researchers argue that oral reading fluency instruments simply measure a student's ability to call words (Pressley, Hilden, & Shankland, 2005; Riedel, 2007; Samuels, 2007). Thus, such a measure of oral reading fluency cannot portray a comprehensive picture of a student's reading ability.

Rasinski and Padak (2004) argue that fluency is the bridge between word recognition and comprehension. The importance of developing students' reading fluency skills is evident in this relationship: fluency is a necessary ability required to develop the ultimate reading skill, which is

to create meaning from text (Bell & McCallum, 2008). Similarly, Daly, Chafouleas, and Skinner (2005) emphasize that developing reading fluency skills is an essential step "to becoming a competent reader, because it increases the student's capacity to use reading as a helpful tool […] with more difficult tasks" (pg. 73). Therefore, in order to obtain a comprehensive picture of a student's reading ability, some educators may find it beneficial to administer an instrument that includes measures of both number of words read correctly within a specified time period as well as a student's understanding of the passage. However, it is important to note that, in some cases, reading fluency predicts students' proficiency very well. Shinn, Good, Knutson, and Tilly (1992) indicate that, in younger students, fluency is the best indicator of reading proficiency. Because measures of oral reading fluency are relatively easy and efficient to administer, rather than administering both fluency and comprehension measures, some educators may prefer to administer fluency measures for the purposes of RTI.

Most CBM measures have an element of time (e.g., number of words read correctly in 1 minute, number of ideas identified correctly within 3 minutes) in common. Researchers have investigated this commonality to determine the effect that time, conceptualized as rate or speed, has on the student's overall performance (Neddenriep, Skinner, Hale, Oliver, & Winn, 2007; Skinner, Neddenriep, Bradley-Klug & Ziemann, 2002; Skinner, Williams, Morrow, Hale, Neddenriep, & Hawkins, 2009; Williams, Skinner, Floyd, Hale, Neddenriep, & Kirk, 2011; Hale, Skinner, Wilhoit, Ciancio, & Morrow, in press). Based on data from these studies, the authors concluded that reading speed is highly correlated to performance on standardized measures (i.e., Woodcock-Johnson Test of Achievement, Broad Reading Cluster, and TCAP reading composite) and is the strongest predictor of overall reading ability.

Of the popular measurement systems currently available (i.e., DIBELS and AIMSweb©), both provide an estimate of a student's reading fluency skills within an oral reading fluency subtest. However, the method chosen by each instrument to measure reading comprehension varies (i.e., DIBELS Retell Fluency subtest, AIMSweb© Maze procedures). For the DIBELS Retell Fluency subtest, in order to assess the student's knowledge of what he/she has recently read, the examiner asks the student to repeat back as much as he/she can remember from the story he/she has just completed reading within a one-minute time period. Students are required to elaborate without the story in their view. This measure is more subjective in its implementation than the oral reading fluency measure. Students may be allowed to make tangential comments, as long as these comments relate, in some way, to the story.

Because of the individual-administration format of DIBELS subtests and the cost associated with teacher training, assessment materials, and data collection of DIBELS subtest scores, Allington (2009) questioned the practicality of using limited resources (e.g., time and money) on a measure that may not accurately measure reading comprehension. Both DIBELS and AIMSweb© have come under the scrutiny of researchers who argue that these assessment systems, as typically implemented within an RTI paradigm, assess only word-calling, rather than comprehension (Pressley et al., 2005; Riedel, 2007; Samuels, 2007). Specifically, Pressley et al. (2005) investigated the predictive utility of the DIBELS Oral Reading Fluency and Retell Fluency subtests to student performance on the TerraNova reading assessment; the authors found that less than 20% of the variance in student scores on the TerraNova reading assessment was accounted for by student performance on the DIBELS subtests. Similarly, Riedel (2007) used first-grade students' scores on the DIBELS Oral Reading Fluency and Retell Fluency subtests to predict end-of-year performance on the Group Reading Assessment and Diagnostic Evaluation.

Results of the Riedel (2007) study indicate that the Oral Reading Fluency subtest was a strong predictor of reading comprehension, as measured by the Group Reading Assessment and Diagnostic Evaluation, correctly identifying student performance on the criterion measure for over 70% of the student sample. Although the DIBELS Oral Reading Fluency subtest was shown to accurately predict reading comprehension on the Group Reading Assessment and Diagnostic Evaluation, Riedel (2007) cautioned against the use of any of the other DIBELS subtests in predicting performance on reading comprehension measures. He indicated that taking into consideration student scores on any of the other subtests did little to inform prediction of student scores on the criterion measure.

In response to Riedel (2007), Samuels (2007) questioned the widescale use of DIBELS subtests in mainstream education. Samuels (2007) proposes that the "simultaneity of decoding and comprehension [. . .] is the essential characteristic of reading fluency" (p. 564). Arguing that fluency involves the ability to correctly and rapidly identify words as well as the ability to comprehend what is being read, Samuels (2007) contends that the DIBELS subtests are more measures of speed than they are measures of fluency. Instead of measuring speed, Samuels indicates that what is needed are tests that require students to decode and comprehend simultaneously, which may occur with contextual reading. However, as mentioned above, assessment within this particular strategy has also been criticized. One new measure that has been developed to address some of the criticisms is available (i.e., the Monitoring Instructional Responsiveness: Reading assessment system), but it is an experimental instrument and empirical support for its psychometric utility is needed.

Utility of the Monitoring Instructional Responsiveness: Reading

Because of the potential problems identified with using the currently-available CBM measures within an RTI paradigm, researchers and educators have begun to create new CBM assessment systems. Because of the limited number of CBM reading assessments that measure both reading fluency and comprehension, one group of researchers has created a CBM reading assessment system, the Monitoring Instructional Responsiveness: Reading (MIR:R; Bell, Hilton-Prillhart, McCallum, & Hopkins, 2012). This particular CBM reading assessment strategy is intended to bridge the assessment gap between reading fluency and comprehension by relying on scores obtained from reading connected text, but the research exploring its utility is limited.

In the first grade MIR:R format, students are required to read short passages (e.g., three sentences or less). The words comprising these sentences have all spacing removed (e.g., thecatisblack). Students are required to slash between words, using a word-chaining technique similar to the TOSCRF (Hammill et al., 2006) and TOSWRF (Mather et al., 2004). The format of the MIR:R for grades two through five is unique. In these grades, students are required to read passages comprised of ten sentences; passages alternate between expository and narrative text. All capitalization and end marks (e.g., periods, exclamation marks, or question marks) have been removed. Students are required to slash between ideas, operationalized as sentences, to correctly identify ten ideas within each passage. Although the word-chaining technique has been adopted elsewhere (i.e., TOSCRF and TOSWRF), the notion of slashing between ideas is unique to the current measure. To ensure adequate difficulty level, all passages for grades 1-5 are written at an end-of-the-year Spache readability level (i.e., 1.9, 2.9, 3.9, etc.).

Slashing between words or between ideas among connected text requires comprehension. Examinees must understand meaningfulness of the content in order to make an appropriate slash

mark. Students' comprehension scores reflect the number of correct and incorrect slashes made for first grade; in higher grades, students' comprehension scores reflect the percentage of ideas correctly identified. Thus, MIR:R scoring incorporates comprehension. MIR:R scoring also incorporates reading fluency. Students are given 3 minutes to complete each CBM probe. The total number words for each probe are tallied alongside each passage, so examiners can easily determine the number of words the student has correctly read, based upon the final slash mark the student has made on the probe. Once the total number of words read has been identified, the examiner divides this number by three to determine the number of words read correctly in 1 minute. Because MIR:R utilizes a group-administration format, it is very time efficient. Additionally, because it incorporates both a reading fluency and reading comprehension measure within one assessment, it is very practical for administration to *all* students, rather than a select few. For the purposes of this study, fluency is defined as the number of words the student has read silently within a 3-minute time period; comprehension is defined as the number of ideas the student has correctly identified within a passage consisting of ideas comprised of connected text in the same 3-minute time period. These particular conceptualizations are relevant for the MIR:R format for grades 2-5. However, reading comprehension is a construct that cannot be directly measured; it must always be indirectly measured. In fact, Pearson (2011) argued that it is impossible to "see" (or directly measure) comprehension. Thus, consistent with previous research (Neddenriep et al., 2007; Skinner et al., 2002; Skinner, et al., 2009; Williams et al., 2011; Hale et al., in press), MIR:R is a measure of reading rate and an indirect measure of reading comprehension.

STATEMENT OF THE PROBLEM

Because of the current emphasis on accountability, teachers require a method of assessment that will assist them in monitoring progress and in predicting student performance on end-of-the-year state achievement tests. Currently, there is some support for use of CBM-type measures for predicting academic success (Crawford et al., 2001; Shapiro et al., 2006; Silberglitt & Hintze, 2005). However, limited evidence is available to support the predictive value of MIR:R, a newly-developed CBM instrument, and no evidence to support its ability to predict high-stakes, end-of-year test scores. In addition, the limited predictive validity data that are available have relied only on initial composite scores and not slope data. For example, initial predictive validity data, using a step-wise regression, indicate that static MIR:R scores predicted 37% ($r^2 = .37$) of the variance in end-of-year STAR Reading Assessment (Renaissance Learning Systems, 1997) scores (Hilton-Prillhart, 2011). Because MIR:R combines reading fluency and reading comprehension using an idea-chaining technique embedded in connected text to create static and slope data for monitoring purposes, it is unique. In order for teachers to be able to predict student performance on end-of-the-year state achievement tests using such measures as MIR:R, it is necessary to establish the relative predictive power of all possible MIR:R scores, not just for static data, but also for slope indices. Consequently, the purpose of this study is to continue the process of establishing the predictive validity of the instrument by examining the relative predictive power of all its scores, including slopes. Specific research questions to be considered follow.

1) To what extent does one static MIR:R Comprehension Rate score predict the TCAP reading composite score?

2) To what extent does one static MIR:R Comprehension Rate score predict TCAP reading composite non-proficiency status?
3) To what extent does the total slope of the MIR:R Comprehension Rate score predict the TCAP reading composite score?
4) To what extent does the total slope of the MIR:R Comprehension Rate score predict TCAP reading composite non-proficiency status?
5) Which MIR:R component score (Comprehension Percentage or Total Words Read) most accurately predicts the TCAP reading composite?
6) Which MIR:R component score (Comprehension Percentage or Total Words Read) most accurately predicts TCAP reading composite non-proficiency status?
7) Which MIR:R score (Comprehension Percentage, Total Words Read, or Slope) most accurately predicts the TCAP reading composite score?
8) Which MIR:R score (Comprehension Percentage, Total Words Read, or Slope) most accurately predicts TCAP reading composite non-proficiency status?

CHAPTER II

METHOD

Participants

Participants for this study included 494 third-grade students from one school district within East Tennessee. The school district from which the participants in this study came is comprised of eight elementary schools. Of the district's student population, 59% are categorized as economically disadvantaged; 47% are male and 53% are female. Approximately 95% of the student population in this particular district is Caucasian.

Instruments

Monitoring Instructional Responsiveness: Reading. Monitoring Instructional Responsiveness: Reading (MIR:R; Bell, Hilton-Prillhart, McCallum, & Hopkins, 2012) is a silent, group-administered, progress-monitoring reading assessment system. Administration time is 3 minutes. For grade three, MIR:R consists of four passages, each containing ten sentences, written at an end-of-grade readability level (i.e., grade level 3.9). The Spache readability formula was selected to reflect text difficulty, or the frequency of words that are considered to be highly decodable and those words that are not as easily decoded (Hilton-Prillhart, 2011; Heibert, 2000). Hilton-Prillhart (2011) chose the Spache readability formula based upon a technical analysis conducted by Good and Kaminski (2001), who characterized this particular readability formula as reliable for texts written at the lower-elementary level. Passages alternate between expository and narrative text and include content and vocabulary from state science and social studies standards for third-grade curriculum. To complete the assessment, students are required to read sentences with no end punctuation marks (e.g., periods, exclamation points, question marks) and without capital letters to signify the beginning of the next sentence. While reading, students must

determine where one idea ends and another begins, then make a slash mark in this place. Slashing between connected text in order to indicate where the end of a sentence or idea would occur is unique to MIR:R.

Another feature of the MIR:R assessment system is that it yields information concerning both reading fluency and reading comprehension. Possible scores for second- through fifth-grade probes include a Total Words Read score as well as a Comprehension Percentage score. The Total Words Read score can be divided by three to indicate the number of words read correctly per minute. The Comprehension Percentage score is derived from the number of ideas a student correctly identified (i.e., made the correct slash mark before the idea began and made the correct slash mark after the idea ended) divided by the number of ideas a student attempted to identify and then multiplied by 100. The Comprehension Percentage score can also be divided by three to indicate the number of ideas identified correctly per minute. These two scores (Total Words Read and Comprehension Percentage) can be multiplied to create a Comprehension Rate score, which indicates an amalgam of number of words read silently and a student's ability to correctly identify a certain number of ideas within a specified time period (i.e., within 3 minutes).

Results from the first year of implementation were used to generate reliability and validity data. Correlation coefficients have been used to determine the psychometric integrity of the MIR:R. For context, according to Sattler (2008), the strength of correlation coefficients can be characterized in the following manner: .20-.29 = low; .30-.49 = moderately low; .50-.69 = moderate; .70-.79 moderately high; .80-.99 = high. For students within the district in which the probes were piloted, alternate-form correlation coefficients ranged from .61-.85, with an average of .75 ($p < .001$). When correlation coefficients were calculated for adjacent probes, the average reliability was found to be slightly higher (.80, $p < .001$). This average reliability coefficient

among adjacent probes is higher than the reliability coefficient calculated between the first and the last probe, which is .69 ($p < .001$), presumably because student performance would be more similar on probes taken closer in time, i.e., minimizing the effects of maturation. Alternate-form reliabilities for all probes are displayed in Table 1.

Concurrent validity estimates between MIR:R and AIMSweb© Maze ranged from .43-.55; the concurrent validity estimate between MIR:R and the STAR test was .67 (Hilton-Prillhart, 2011). Hilton-Prillhart (2011) compared the predictive utility of MIR:R and AIMSweb© Maze scores to estimate end-of-year STAR scores and, using a step-wise multiple regression, found that MIR:R scores predicted 37% of the variance in the STAR scores and was the most powerful predictor; AIMSweb© scores failed to produce additional predictive variance. Other reliability and validity data for grades 1 through 3 may be accessed from Hilton-Prillhart (2011).

Tennessee Comprehensive Assessment Program (TCAP) Achievement Test. The Tennessee Comprehensive Assessment Program (TCAP) Achievement Test is an English-only, criterion-referenced, standardized test (Tennessee Department of Education, http://tn.gov/education/assessment/achievement.shtml). The test is administered each spring to students in grades 3-8, as required by the State of Tennessee Department of Education. The test includes a multiple-choice format across five different content areas: Reading, Language Arts, Mathematics, Science, and Social Studies. The TCAP is also timed. Internal consistency reliability coefficients are reported as ranging from .95 to .96 (Miller, DeLapp, & Driscoll, 2007). Additionally, the concurrent validity coefficients resulting from the analysis of TCAP and AIMSweb© Oral Reading and Maze subtests reported by Yeo et al. (2012) range from .51 to .75.

Scores provided by the TCAP include raw scores and scaled scores for each content area composite. The State of Tennessee Department of Education has established cut scores in order

to categorize student performance according to level of proficiency: below basic, basic, proficient, and advanced. Scores of below basic, basic, proficient, or advanced on the TCAP reading composite are equivalent to a Reporting Categories Performance Index (RCPI) score of 38 or lower, 39-69, 70-87, and 88 or better, respectively; 69 or lower is considered to represent *non-proficiency.* The RCPI score is an estimate of the number of items a student would be expected to answer correctly if there had been 100 items within that category. There are a total of 39 items for the TCAP reading composite. For the purposes of this study, the Reading/Language Arts composite score was used.

In order to determine whether students are making adequate yearly progress, as required by NCLB, state departments of education and local education authorities may establish cut scores defining various levels of achievement. In Tennessee, for the year of 2010, before any data are disaggregated, the percentages for all students were: 12.3%, 37%, 38.3%, and 12.4%, for the below basic, basic, proficient, and advanced achievement groups, respectively (Tennessee Department of Education, http://tn.gov/education/assessment/achievement.shtml). The TCAP results for the school system from which this sample was drawn did not align with the percentages outlined by the State Department of Education. Specifically, the results obtained by this sample were: 10.4%, 50.1%, 32.2%, and 7.2%. The system had more students who performed at the basic level than did the entire state based on the State's guidelines used to determine adequate yearly progress. When the achievement groups are combined to create proficiency status, approximately 61% of the scores from this sample are classified as non-proficient (i.e., below basic and basic) and 39% are classified as proficient (i.e., proficient and advanced), as compared to 49.3% and 50.7%, respectively, for the entire state.

Procedures

The following procedures were those used by researchers in the initial study in which the MIR:R probes were piloted; the MIR:R data used in the current study are data that were collected during the pilot study (for detailed information, see Hilton-Prillhart, 2011). The TCAP data were collected by teachers using the standardization procedures outlined within the TCAP manual during the spring of the year in which the pilot data were collected (Fall 2009 - Spring 2010).

For pilot data collections, researchers obtained permission to conduct the study from the district's Superintendant, building-level administrators, and the University's Institutional Review Board. Next, teachers and district-level reading specialists were trained in the administration of the MIR:R probes and practice sheets. Classroom teachers then administered the MIR:R probes approximately every two weeks for a total of ten administrations, beginning in November of 2009. District personnel completed the scoring of the probes and data entry of students' scores.

Administration. MIR:R probes were administered in a group format to third-grade students in their regular classrooms by their classroom teacher. Before administering the probes, teachers provided an opportunity to practice the probe format on a designated practice sheet, which contained opportunities for assisted and independent practice. Teachers were provided with, and read, scripted practice instructions to further illustrate the slashing procedure required by the MIR:R probes. Following completion of the practice sheets, teachers then administered the MIR:R probes.

After students wrote their names and dates on their MIR:R probe sheets, teachers read the scripted administration directions. Teachers then clarified any student questions and began timing the probe administration to ensure the probes were administered as intended. Students were instructed to work through the passages for 3 minutes. At the end of 3 minutes, teachers

instructed students to stop their work on the MIR:R probe sheets, which were then collected by the teacher. This same procedure was followed for each administration. TCAP achievement tests were administered and proctored by classroom teachers in the spring of the academic year, according to the Tennessee State Department of Education's administration guidelines.

Scoring. District personnel completed the scoring and data entry of all student probes. Scores produced by the MIR:R probes include an Ideas Attempted score, an Ideas Correct score, and a Total Words Read score. To produce a Comprehension Percentage score, the Ideas Correct score was divided by the Ideas Attempted score. In order to determine the Total Words Read score, word counts were provided in a column to the right of the passages on each probe sheet to allow for easy calculation of the total number of words read by each student. The word before the students' last slash on the page was counted as the last word read by the student. To determine the number of ideas the student attempted to identify (i.e., make slashes before/after the identified idea), scorers located the last slash mark the student made. After identifying this slash, scorers identified the idea number (e.g., 1-40) in a column that tallied the number of ideas in each passage for all four passages that corresponded to the student's last slash. Finally, to determine the number of ideas correctly identified by a student, scorers looked for slash marks that offset each individual idea within the passage. For instance, in the second idea/sentence of the passage, there should be one slash before the beginning of the second idea (that identified the end of the first idea) and one slash at the end of the second idea. There should be no slash(es) made within the idea in order for the idea to count as correct. Because scoring allows for a Total Words Read score and a Comprehension Percentage score, MIR:R probes take a student's skills in both fluency and comprehension into consideration. Scoring accuracy was checked by a team of researchers who compared 65% of the hard-copy protocols with the original scores entered

into the system's database to ensure that the scoring/entering error percentages did not exceed 5%. TCAP achievement tests were scored according to the Tennessee State Department of Education's scoring guidelines.

Data Cleaning

Because of wide variability in student scores, extreme outliers were removed from the data set. Identification of outliers was determined by scores greater than three standard deviations above the mean. In total, 9 outlying student scores were removed from the data set. With these outlying scores removed, the number of students in the sample decreased to 485.

Slope Determination

Several slope indices created by collecting probes less frequently than the slope obtained from all the probes were compared. The slope indices were calculated via latent growth curve analyses. That is, the slope index comprised of all probe data was first determined for each student (i.e., data points were collected over all ten administrations). Then, an additional slope for each student was created by including every other probe. Two slopes were created using the every-other-probe criteria (i.e., one slope was created from all even data points, then a second from all odd data points). Finally, a slope was created for each student using every third data point (i.e., first, fourth, seventh, tenth data points). To allow for comparisons among slopes, Slopes 2, 3, and 4 were adjusted (i.e., divided by either 2 or 3, depending on the number of weeks between included data points). The mean of Slope 1, comprised of every data point obtained from the every-2-week monitoring schedule, is 3.51 (range = -20.26 – 21.62; *SD* = 6.10). The mean of Slope 2, comprised of the even data points obtained from the every-4-week monitoring schedule, is 4.33 (range = -24.42 – 31.58; *SD* = 7.63). The mean of Slope 3, comprised of the odd data points on the same every-4-week schedule, is 2.79 (range = -19.13 –

25.39; *SD* = 7.47). The mean from Slope 4, created from using every third data point obtained on the every-6-week monitoring schedule, is 4.18 (range = -15.95 – 23.46; *SD* = 7.01). Descriptive statistics for these slopes are displayed in Table 2.

Three separate *t* tests were conducted to determine whether the values obtained for the slopes were significantly different. These analyses resulted in significant differences between the values obtained for all of the slopes. When comparing Slope 1 and Slope 2, the *t* value obtained is -3.94 ($p < .001$). When comparing Slope 1 and Slope 3, the *t* value obtained is 3.61 ($p < .001$). Finally, when comparing Slope 1 and Slope 4, the *t* value obtained is -3.55 ($p < .001$). *T* test statistics are displayed in Table 2. Variations in time between administrations (i.e., between 2 and 4 weeks, due to school breaks and TCAP testing) may have influenced the differences among slopes, impacting the results of the *t* tests.

Pearson product-moment correlation coefficients were also calculated to determine relationships between the various slopes. Correlation coefficients ranged from .77 (Slopes 1 and 3; $p < .01$) to .83 (slopes 1 and 4; $p < .01$). Slope 4 (i.e., the slope comprised of every third data point or the every-6-week monitoring schedule) is most closely correlated to slope 1 (i.e., the slope comprised of every data point or the every-2-week monitoring schedule), with a correlation coefficient of .83 ($p < .01$).

Coefficients of determination, r^2, can be calculated to illustrate the amount of shared variance between two variables. The coefficients of determination were calculated between the four slopes to determine the strength of the relationships among the various slopes. The coefficients of determination ranged from .59 (Slopes 1 and 3) to .69 (Slopes 1 and 4). Although any of the slopes may be reasonable to use for monitoring, Slope 1 was chosen for the data analyses for this study because it reflects the entire data set. These same study/research questions

may be explored using Slopes 2, 3, or 4; exploring these questions should be the focus of future research.

Analyses

Student scores across all MIR:R probes and student scores from the Spring 2010 TCAP reading composite were used for various data analyses. Because slopes take into account students' growth rates over time, rather than a representation of a student's performance at a single point in time, Comprehension Rate slopes were used as a predictor variable and compared to static MIR:R scores. The static MIR:R scores used were those collected during the sixth probe administration. This particular administration was chosen in order to allow students and teachers enough time and exposure to adapt to the unique format of MIR:R, but was still somewhat distant (i.e., 6 – 8 weeks) from the criterion measure.

To investigate the relationship between the students' MIR:R Comprehension Rate static score and slope score to predict TCAP performance, a Pearson product-moment correlation coefficient was calculated. The relationship between the MIR:R Comprehension Rate static score and slope score and a score of *non-proficient* (vs. proficient) on the TCAP was investigated using a point biserial correlation coefficient. In order to determine which MIR:R component score (Comprehension Percentage or Total Words Read) best predicted TCAP performance, stepwise multiple regression was used. Logistic regression was used to determine which MIR:R component score (Comprehension Percentage or Total Words Read) best predicted TCAP non-proficiency status, a categorical dependent measure. Stepwise multiple regression was employed to compare the relative predictive power of three MIR:R scores (Comprehension Percentage, Total Words Read, and Comprehension Rate Slope) to predict TCAP performance; logistic regression was used to determine which (of three) MIR:R component scores (Comprehension

Percentage, Total Words Read) and composite score (Comprehension Rate Slope) best predicted TCAP non-proficiency status. Finally, sensitivity and specificity data were calculated to determine the accuracy of non-proficiency predictions based upon MIR:R scores. Each research question is addressed and the relevant data are presented in the following sections.

CHAPTER III

RESULTS

To examine the predictive validity of the Monitoring Instructional Responsiveness: Reading (MIR:R; Bell, Hilton-Prillhart, McCallum, & Hopkins, 2012), data were analyzed for a universal screener and nine probes, administered to third-grade students receiving both general- and special-education instruction. Multiple MIR:R scores, both static and slope, were used to investigate the relative predictive power of the MIR:R assessment system to predict end-of-the-year reading achievement performance on the Tennessee Comprehensive Assessment Program, (TCAP) Achievement Test (Tennessee Department of Education, http://tn.gov/education/assessment/achievement.shtml). The MIR:R slope scores reflected all probes' data points for each student; the static scores were those collected during the sixth probe administration.

Use of TCAP as Criterion Score

For the purposes of this study, the TCAP reading composite score was chosen as the criterion against which all scores were compared. For some analyses, the specific reading composite score was used. For other analyses, proficiency status was determined as follows. Student performance on the TCAP reading composite was categorized according to the levels of proficiency associated with the cut scores established by the State of Tennessee Department of Education (e.g., below basic, basic, proficient, and advanced). These categories can be combined to represent non-proficiency (e.g., below basic and basic categories) and proficiency (e.g., proficient and advanced categories). Results of the TCAP reading composite indicate that approximately 61% of the students' scores fell within the non-proficiency category; 39% of the students' scores fell within the proficiency category.

Descriptive Statistics of MIR:R and TCAP Scores

Descriptive statistics were calculated for each of the static and slope scores and for the TCAP composite. Of the MIR:R component scores, the mean Comprehension Percentage score is 50.73 (range = .00 – 100.00; *SD* = 30.70); the mean Total Words Read score is 250.61 (range = 4.00 – 412.00; *SD* = 86.01). The mean Comprehension Rate total score from the sixth probe administration is 124.91 (range = .00 – 412.00; *SD* = 86.31). Finally, the mean Comprehension Rate slope score, comprised of data points from all administrations, is 3.51 (range = -20.26 – 21.62; *SD* = 6.10). The mean TCAP reading composite scale score is 749.17 (range = 600 – 857; *SD* = 33.20). These descriptive statistics can be found in Table 3. Additionally, correlation coefficients were calculated among all possible scores. These coefficients ranged from -.16 (Total Words Read and Comprehension Percentage; $p < .01$) to .60 (TCAP reading composite score and Comprehension Rate static score; $p < .01$). All correlation coefficients are shown in Table 4. Of all relationships expressed, the most significant is that between the TCAP reading composite score and the Comprehension Rate static score, which is illustrated by the highest correlation coefficient obtained.

Predictive Validity of the MIR:R Overall Static Score

The relationship between the MIR:R Comprehension Rate static score and the TCAP reading composite was explored by calculating a Pearson product-moment correlation coefficient. This particular analysis resulted in a coefficient of .60 ($p < .01$). According to Wayman, Wallace, Wiley, Ticha, and Espin (2007), coefficients of .70 and above, .50-.69; and .49 and below may be characterized as high, moderate, and low, respectively. Using these characterizations as a guideline, the coefficient of .60 represents a moderate relationship between the MIR:R Comprehension Rate static score and the TCAP reading composite. Another way to

conceptualize the amount of shared variance between the MIR:R Comprehension Rate static score and the TCAP reading composite is to calculate a coefficient of determination (r^2). The coefficient of determination for this particular relationship is .36, indicating that the MIR:R Comprehension Rate static score explains approximately 36% of the variance in the TCAP reading composite scores.

The relationship between the MIR:R Comprehension Rate static score and non-proficiency status on the TCAP reading composite was determined via calculation of a point biserial correlation coefficient. This analysis resulted in a correlation coefficient of .52 ($p < .01$). This particular coefficient also represents a moderate relationship between the MIR:R Comprehension Rate static score and TCAP non-proficiency status. Receiver operating characteristic (ROC) curves provide a useful strategy for interpreting how accurately a measure can predict a categorical variable (Johnson et al., 2009). Specifically, ROC curves offer an overall indication of diagnostic accuracy via examination of the area under the curve (AUC). AUC values close to 1 indicate strong prediction; AUC values close to .5 indicate prediction approximating chance (Johnson, et al., 2009; Zhou, Obuchowski, & Obushcowski, 2002). The ROC curve value calculated to determine how accurately the MIR:R Comprehension Rate static score predicted scores within the non-proficient range on the TCAP reading composite resulted in a value of .81, indicating that the Comprehension Rate static score offers a moderately strong ability to predict non-proficient scores, according to the classification guidelines offered by Johnson et al. (2009) and Zhou et al. (2002).

The ROC analysis also provides a method to determine the accuracy of the students' predicted proficiency categorization on the TCAP reading composite by calculating sensitivity and specificity levels. Sensitivity and specificity levels describe how accurately the test

discriminates between students who are likely to obtain a score of non-proficiency or proficiency on the TCAP reading composite. Sensitivity is calculated as the ratio of true positives (students correctly identified as at risk for scoring within the non-proficient range who later score within the non-proficient range) relative to the sum of true positives and false negatives (students who are not identified as at risk for scoring within the non-proficient but later score within the non-proficient range). Specificity is calculated as the ratio of true negatives (students correctly identified as proficient who later score within the proficient range) relative to the sum of true negatives and false positives (students who are not identified as proficient but later score within the proficient range). It is possible to set a cut score associated with certain levels of either sensitivity or specificity; then, one can determine the level of specificity or sensitivity associated with that particular cut score. In order to capture the majority of the students whose scores fall within the non-proficient range, sensitivity levels were set to 90%. For the analysis investigating the MIR:R Comprehension Rate static score and TCAP non-proficiency status, the corresponding cut score closest to 90% is identified as 175.55; the corresponding specificity level is 55%. True and false positive identifications are 227 and 69, respectively; true and false negative identifications are 110 and 43, respectively. These ratios lead to a classification accuracy of 75%. Detailed sensitivity and specificity information for this particular analysis is shown in Table 5.

Predictive Validity of the MIR:R Overall Slope Score

The relationship between the MIR:R Comprehension Rate slope and student performance on the TCAP reading composite was determined by calculating a Pearson product-moment correlation coefficient. This analysis resulted in a coefficient of .22 ($p < .01$). The obtained

coefficient represents a weak relationship between MIR:R Comprehension Rate slope and the TCAP reading composite.

In addition, the relationship between the Comprehension Rate slope and non-proficiency status on the TCAP reading composite was determined via calculation of a point biserial correlation coefficient. This analysis resulted in a coefficient of .20 ($p < .01$). The obtained coefficient represents a weak (but significant) relationship between the MIR:R Comprehension Rate slope and TCAP non-proficiency status. A ROC curve was calculated, with an accompanying AUC value, to determine how accurately the MIR:R Comprehension Rate slope predicted scores within the non-proficient range on the TCAP reading composite. The AUC value obtained via this analysis is .62, indicating some gain over chance in accurately predicting scores within the non-proficient range on the TCAP reading composite using the MIR:R Comprehension Rate slope score. For this analysis, when the sensitivity level is set at 90%, the corresponding cut score closest to 90% is identified as 9.65; the corresponding specificity level is 24%. True and false positive identifications are 254 and 140, respectively; true and false negative identifications are 45 and 28, respectively. These ratios lead to a classification accuracy of 64%. Detailed sensitivity and specificity information for this analysis may be found in Table 6.

Relative Predictive Power of the MIR:R Component Scores

The relationship between the MIR:R component scores (Comprehension Percentage and Total Words Read) and student performance on the TCAP reading composite was determined via a step-wise multiple regression. Both component scores provide predictive utility. The MIR:R Comprehension Percentage score entered the equation first and predicted 9.4% of the variance in the TCAP scores ($R^2 = .09$; $p < .01$); the MIR:R Total Words Read score predicted an

additional 1.1% of the variance in the TCAP scores ($R^2 = .01$; $p < .05$). Detailed results of this particular step-wise multiple regression analysis can be found in Table 7.

When all three MIR:R scores (Comprehension Percentage, Total Words Read, and Comprehension Rate static score) were entered into a multiple regression, the combined scores predicted 37.6% of the variance in TCAP scores ($R^2 = .38$; $p < .01$). When you add together the predictive utility of the two component scores (Comprehension Percentage and Total Words Read scores), approximately 10.5% of the variance in TCAP scores is predicted ($R^2 = .11$; $p < .05$). Thus, the Comprehension Rate static score independently predicts approximately 27.1% of the variance in TCAP scores ($R^2 = .27$; $p < .05$). Although the two component scores are combined to form the Comprehension Rate static score (Comprehension Rate = Comprehension Percentage x Total Words Read), the component scores are calculated independently (Comprehension Percentage = number of ideas correctly identified divided by the number of ideas the student attempted to identify; Total Words Read = the number of words the student read within the 3-minute time period). Consequently, the Comprehension Rate score offers the most predictive utility when predicting student performance on the TCAP reading composite.

The relationship between the component scores and TCAP reading composite non-proficiency status was determined via logistic regression. The MIR:R Comprehension Percentage score accurately predicted 84.1% of non-proficient scores and accurately predicted 60.9% of the proficient scores on the TCAP reading composite ($p < .01$). The MIR:R Total Words Read score accurately predicted 87.8% of the non-proficient scores and accurately predicted 59.8% of proficient scores on the TCAP reading composite ($p < .01$). ROC AUC values were calculated, also. The AUC value calculated from the Comprehension Percentage score is .80, indicating a moderately strong ability to accurately predict scores within the non-

proficient range on the TCAP reading composite. The AUC value calculated when using the Total Words Read score to predict scores within the non-proficient range on the TCAP reading composite is .55, indicating that MIR:R Total Words Read predicts TCAP non-proficiency at about a chance level.

Finally, sensitivity and specificity indices were calculated for both Comprehension Percentage and Total Words Read. When the sensitivity level is set at 90% for the Comprehension Percentage component score, the corresponding cut score closest to 90% is identified as 74.54; the corresponding specificity level is 52%. True and false positive identifications are 228 and 70, respectively; true and false negative identifications are 109 and 43, respectively. These ratios lead to a classification accuracy of 75%. When the sensitivity level is set at 90% for the Total Words Read component score, the corresponding cut score closest to 90% is identified as 396.50; the corresponding specificity level is .06%. True and false positive identifications are 238 and 72, respectively; true and false negative identifications are 107 and 33, respectively. These ratios lead to a classification accuracy of 77%. Detailed sensitivity and specificity information for these analyses may be found in Table 8.

Relative Predictive Power of MIR:R Component Scores and Slope

The ability of the component scores (Comprehension Percentage and Total Words Read) and slope to predict student performance on the TCAP reading composite was determined using a step-wise multiple regression analysis. Both of the static component scores and the slope score offer some predictive utility. The Comprehension Percentage score is the most powerful, predicting 9.3% of the variance in TCAP scores ($R^2 = .09$; $p < .01$). The Comprehension Rate slope predicts an additional 2.4% of the variance in TCAP scores ($R^2 = .02$; $p < .01$); the Total

Words Read score predicts an additional .8% of the variance in TCAP scores ($R^2 = .01$; $p < .05$). Detailed results of this step-wise multiple regression may be found in Table 9.

Finally, component scores and slope values were included in a logistic regression to determine the extent to which each of these MIR:R indices predict scores within the non-proficient range on the TCAP reading composite. Although the results of the multiple regression indicate that the Comprehension Rate slope score did contribute significantly to the prediction of TCAP, the logistic regression found that this particular score did not contribute significantly to the prediction of TCAP; this score was removed from the equation. Consequently, the results of this logistic regression, related ROC AUC values, and sensitivity and specificity analyses remain the same as previous analyses investigating the extent to which the component scores predict TCAP non-proficiency status.

CHAPTER IV

DISCUSSION

Rather than relying on end-of-year scores to determine student progress, educators need more timely indicators of at-risk status for particular students. Early identification of academic problems should lead to more timely and effective interventions. One way to measure progress early (and often) is to use formative assessment methods. Adherence to the Response to Intervention (RTI) paradigm requires that formative assessment be used to determine at-risk status, and, then, whether empirically-validated interventions actually improve student performance on Curriculum-Based Measurement (CBM) tests. Presumably, results of these CBM tests are used to alter or modify the curriculum and/or instruction to better address the needs of students who may not be performing as well as their peers. If instructional changes do not lead to improved outcomes on subsequent CBM tests, students may be considered for special education placement.

CBM Within Response to Intervention Models

The RTI/problem-solving paradigm requires that each student be assessed. There is an emphasis on measuring students' reading skills within this model. There are a variety of instruments available to measure student progress; however, the administration format and aspects of reading that are assessed by each instrument vary. Because of the current emphases on accountability for *all* students and attaining adequate yearly progress, as well as how these goals are ascertained (e.g., performing in a certain manner on end-of-the-year achievement tests), efficiency of test administration and utility of assessment results become paramount issues. The types of information and the amount of information gleaned from just one assessment must be weighed against the time required for administration to all students. Therefore, assessments

should be efficient, yield high-quality, desired information, and be fairly easy to administer and interpret.

Currently-available instruments designed to monitor students' progress toward state-mandated standards measure reading fluency and may measure reading comprehension. The progress-monitoring measures that are currently popular in U.S. school systems (e.g., DIBELS and AIMSweb©) require multiple subtests to be administered in order to obtain information concerning both reading fluency and reading comprehension; further, most of the measures require individual administration. Thus, administration of these subtests to all students requires much time and can be costly (e.g., multiple record forms for each student, volunteer assessment hours, etc.). And, use of these measures is controversial. There is some support in the literature for using oral reading fluency to assess reading (Shinn et al., 1992). However, fluency measures are not without critics. For instance, Allington (2009) questioned the practicality of using limited resources (e.g., time and money) on a measure that only indirectly measures reading comprehension. The empirical support for these measures is mixed. In one study, only the DIBELS Oral Reading Fluency subtest significantly predicted the end-of-year performance on the Group Reading Assessment and Diagnostic Evaluation; the other subtests offered little predictive utility (Riedel, 2007).

Others who question the adoption of DIBELS into mainstream education (Pressley, et al., 2005) argue that the way oral reading fluency is currently operationalized is no more than a word-calling measure. However, it is the "simultaneity of decoding and comprehension [. . .that] is the essential characteristic of reading fluency" (Samuels, 2007, p. 564). According to Samuels (2007), measures that require students to decode and comprehend simultaneously are needed.

The Monitoring Instructional Responsiveness: Reading (MIR:R; Bell, Hilton-Prillhart, McCallum, & Hopkins, 2012) assessment program was developed to address this need.

MIR:R, piloted as part of a comprehensive RTI in a school district within East Tennessee, uses a group-administration format and incorporates both reading fluency and reading comprehension within one measure, making it potentially efficient and practical. Additionally, administration and scoring procedures are fairly simple. MIR:R has the potential to yield useful information in an easily-interpreted format, but validity data are limited. The purpose of this study is to continue the process of establishing validity, and, particularly, to determine its ability to predict high-stakes, end-of-year scores.

Relationships Among the MIR:R Overall Static Score, Slope, and TCAP: Zero-Order Correlational Analyses

Correlation coefficients obtained between the overall MIR:R Comprehension Rate static scores and TCAP performance and non-proficiency status ranged from .52 - .60 ($p < .01$), and indicate moderately strong relationships. Apparently, MIR:R predicts high-stakes, end-of-year scores reasonably well, and its predictive power is comparable to most other CBM-type measures (e.g., DIBELS, AIMSweb, and independently-created measures) in the literature (Crawford et al., 2001; Shapiro et al., 2006; Silberglitt & Hintze, 2005). Specifically, Crawford et al. (2001) compared reading rate performance on passages constructed from Houghton Mifflin Basal Reading Series to the Oregon Department of Education's end-of-year reading achievement test; correlations ranged from .60 - .66 ($p < .01$).

Similarly when Shapiro et al. (2006) compared the utility of both AIMSweb and an independently-created CBM assessment system to predict student performance on the end-of-year measure for the state of Pennsylvania (the Pennsylvania System of School Assessment;

PSSA), they found moderate relationships between both CBM assessment systems and the PSSA, with correlation coefficients ranging from .62 - .69, with one exception – a coefficient of .25, illustrating the relationship between the fall CBM administration and the PSSA ($p < .01$). Finally, when examining correlations between student scores on an independently-created CBM reading assessment system and the Minnesota Comprehensive Assessment (MCA) reading achievement test, Silberglitt and Hintze (2005) found that correlations ranged from .68 - .71 ($p < .01$), depending on whether the fall, winter, or spring CBM score was used to predict student performance on the MCA. This range of correlation coefficients is slightly higher than is typically found in the literature and relative to coefficients found in the current study.

Because results from the current study are consistent with most in the literature, these data provide additional evidence supporting the use of curriculum-based measures to monitor students' progress and, ultimately, to predict performance on an end-of-year, high-stakes achievement measure. Like previous studies, correlations obtained between the MIR:R Comprehension Rate static score and performance on the TCAP reading composite indicate a moderate relationship between the predictor and criterion score.

Although the data showing moderate relationships between CBM-type static scores and end-of-year scores are encouraging, they represent the relationship between two static scores. Some suggest the predictive relationship between slope scores (comprised of gains from probe-to-probe administrations throughout the year) should be even more powerful because they rely on multiple values. However, in this study, correlations obtained between the overall MIR:R Comprehension Rate slope and TCAP performance and non-proficiency status range from .20 - .22 ($p < .01$). Although these coefficients are statistically significant (because of the large sample size), they indicate a weak relationship between the Comprehension Rate slope and the TCAP

scores. When TCAP scores are used to create a dichotomous variable (e.g., non-proficient vs. proficient), the range of possible scores is reduced, which may limit predictive power. However, these slope score results are similar to those reported by Yeo et al. (2012), who found that CBM growth estimates, as calculated via student performance on AIMSweb© oral reading fluency and maze subtests, offered little predictive utility when predicting student performance on the TCAP.

Specifically, Yeo et al. (2012) found that none of the predictions of the oral reading fluency and maze CBM slopes were statistically significant. Growth rates simply did not predict performance on the TCAP in a significant manner. Yeo et al. (2012) concluded that the unstable nature of the slope led to decreased predictive utility; unstable scores may affect MIR:R prediction, as well. Some evidence in support of this hypothesis is available from Hilton-Prillhart (2011). In an early study using generalizability theory analyses, she found that five consecutive probes were required to obtain a reliability coefficient of .80.

Although the predictive capability of the MIR:R and the CBM measures used by Yeo et al. (2012) are similar, the instruments are not. The CBM instruments used in the Yeo et al. (2012) study were either oral reading fluency or reading comprehension (as measured by the maze procedure); MIR:R incorporates aspects of both reading fluency and reading comprehension measurement within one instrument.

Finally, Stage and Jacobsen (2001) investigated the predictive power of an independently-created CBM assessment system to predict student performance on the Washington Assessment of Student Learning (WASL). Stage and Jacobsen (2001) found that correlations between slopes comprised of CBM data and student performance on the WASL ranged from .26 - .35 ($p < .01$). These correlations, similar to those obtained in the current study,

indicate a weak relationship between the CBM slope and student performance on the end-of-year achievement measure. In general, slope measures are not as powerful as static scores.

Other researchers report stronger relationships between slope scores and future performance on standardized measures. For example, Schatschneider et al. (2008) investigated the predictive utility of DIBELS ORF slopes to predict student performance on the Stanford Achievement Test, 10th edition (SAT; Harcourt Brace, 2003), and reported correlation coefficients ranging from .52 to .69 ($p < .0001$), using criterion data collected either at the end of the first or second year of the study. DIBELS ORF slopes were calculated across four assessment periods (e.g., fall, early winter, late winter, spring) using both a linear growth model and a quadratic growth model in year one. A higher correlation coefficient was found between DIBELS ORF slopes and student performance on the SAT during the first year than during the second year. The higher correlations found during the first year may be due to less time in between the administration of the predictor and criterion variables. While the correlations indicate moderate predictive utility, the authors concluded that growth estimates provided by the slopes did not add unique information to the prediction of future reading performance above what could be obtained via a single score. Importantly, although the correlations between the slope and criterion measure resulting from the Schatschneider et al. (2008) study are higher than those found in the current study, the increased ability of a single, static score to predict student performance on an end-of-year measure is similar across both studies.

Relationships Among MIR:R Component Scores, Slope Score, and TCAP: Multivariate Analyses

MIR:R yields two component scores. Which one is most predictive of TCAP? To determine the answer, both scores (i.e., Comprehension Percentage and Total Words Read) were

used to predict TCAP performance and TCAP non-proficiency status. Comprehension Percentage was determined to be a better predictor than Total Words Read. Comprehension Percentage accounted for approximately 9.4% of the variance in the TCAP scores; Total Words Read predicted an additional 1.1% of the variance ($p < .05$). When comparing the accuracy of the proficiency-status predictions made by the two MIR:R component scores, the results of the logistic regression indicate that the MIR:R Comprehension Percentage score provides only slightly more accurate predictive utility. Specifically, the Comprehension Percentage score accurately predicted 84% of the non-proficient TCAP scores, whereas the Total Words Read score accurately predicted 88% of the non-proficient scores ($p < .01$).

When the Comprehension Rate slope was added as a predictor variable to the step-wise multiple regression, the Comprehension Percentage component score was again more powerful, predicting 9.3% of the variance in TCAP scores, with the slope score and Total Words Read component score predicting an additional 2.4% and 0.8% of the variance, respectively ($p < .05$). These relationships are also apparent from the zero-order correlations. The correlation coefficient between Total Words Read and TCAP scores is .06, but is .31 ($p < .01$) between Comprehension Percentage and TCAP scores. When the Comprehension Rate slope was added as a predictor variable to the logistic regression, the slope did not offer any significant additional predictive utility.

Of the two component scores, the Comprehension Percentage score is the most powerful predictor of end-of-year TCAP performance. Timed comprehension (the percentage of ideas correctly identified by the student during the 3-minute time period) may provide most of the predictive power. However, not all studies report data consistent with this result. For example, Neddenriep et al. (2007), Skinner et al. (2002), Skinner et al. (2009), Williams et al. (2011), and

Hale et al. (in press) found that reading speed was the strongest predictor of overall reading ability, and concluded that rate was highly correlated to both reading fluency and reading comprehension performance on standardized achievement measures (i.e., Woodcock-Johnson Tests of Achievement, Broad Reading Cluster, and TCAP reading composite). Even though the results seem contradictory superficially, there is a common element between results showing stronger rate (than comprehension) prediction and the current findings. That is, the CBM rate-based measures are timed, but so is the MIR:R Comprehension Percentage score. It is a rate measure, in a sense.

The finding showing a stronger Comprehension Percentage – TCAP relationship (relative to the Total Words Read score) may be explained by the stronger match between the task of identifying ideas within a passage of connected text and the tasks required by the TCAP reading composite. In any case, these results seem to support Samuels' (2007) admonition to include instruments that measure both reading fluency and reading comprehension within a connected-text format. In fact, results of this study provide verification of Samuels' observation. The single, strongest predictor is the Comprehension Rate static score. It predicts 27.1% ($R^2 = .27$; $p < .05$) unique variance in TCAP reading composite scores. Thus, when both aspects of reading (i.e., fluency and comprehension) are taken into account together, prediction improves. The MIR:R Comprehension Rate static score is unique because it includes both aspects of reading in a timed manner. Thus, this particular score contains elements that experts have found to be good predictors of future reading success.

Accuracy of MIR:R Predictions to TCAP Non-Proficiency Status

Results from the analysis showing relationships between MIR:R component scores and the TCAP reading composite performance are consistent with those showing the relationship

between MIR:R component scores and TCAP non-proficiency status. Johnson et al. (2009) found similar results when they investigated various DIBELS subtests to determine the accuracy of these screening instruments to predict student performance and a dichotomized variable, satisfactory/unsatisfactory performance, on the Florida Comprehensive Assessment Test (FCAT). They found that combining various DIBELS subtests offered small increases in specificity when sensitivity was set to 90%. When comparing a dichotomous classification scheme of their criterion measure (i.e., the FCAT) to the three cut-score levels proposed by DIBELS, Johnson et al. (2009) concluded that using the dichotomous classification scheme appeared to have greater predictive utility, though some students were still incorrectly classified. The classification accuracy evidenced by DIBELS ranged from 75% (using the Nonsense Word Fluency subtest) to 77% (using the Oral Reading Fluency Subtest). In the current study, classification accuracy ranged from 64% (using the MIR:R Comprehension Rate slope) to 77% (using the Total Words Read component score). Apparently, the classification accuracy produced by the MIR:R assessment system is similar to DIBELS.

Summary

Overall, these data provide some evidence of predictive validity of the MIR:R. While the Comprehension Rate slope yielded limited utility to predict TCAP scores, the component scores were stronger. The Comprehension Percentage component score offers greater predictive utility than the Total Words Read component score or any slope measure. Results indicate that the Comprehension Percentage score predicts student performance and non-proficiency status on the TCAP reading composite at a moderate level, while neither the Total Words Read component score nor the slope adds appreciably.

Limitations and Future Research

Several limitations exist within the current study. One potential limitation involves fidelity of both administration and scoring. Although administration guidelines and checklists were given to the teachers during the initial training, it was not possible for researchers to observe every MIR:R administration. Variation in the amount of time between probe administrations is another limitation; this variation may have impacted the results of analyses investigating the differences between slopes. Additionally, although a team of researchers checked overall scoring and data entry, this included only a sample of the student probes. It is possible that scoring fidelity was compromised in probes that were not checked. Finally, the amount of time between the MIR:R administration from which all static scores were collected (e.g., the sixth administration) and the TCAP administration was brief (i.e., 6-8 weeks).

Another limitation is that treatment information is limited. Although students receiving both general-education and special-education instruction were included in the population, these students were not distinguished in the overall data set. Additionally, the sample is limited in terms of both ethnicity and region. Educators may be interested to learn how well the MIR:R assessment system predicts performance on an end-of-year measure for more diverse groups of students, especially as data disaggregation by group is required by NCLB.

Future research may extend this assessment approach to passages comprised of connected text, both within passages and between passages. Because MIR:R passages are comprised of both expository and narrative text within the same probe, it is possible that this format affected student performance on the measure. For instance, the first passage may be comprised of expository text that includes vocabulary from state science standards, whereas the second passage may be comprised of narrative text about fishing from a pier on the beach. Because the

text is not connected from one passage to the next, students are required to change their reading points of reference every ten sentences. Higher-performing students may not struggle with changing their reading points of reference; however, lower-performing students may struggle with this change, which may impede their overall performance. Future research should also investigate the utility of this assessment approach using text that is connected from one passage to the next, with latter passages building upon content in former passages. Future research should also investigate the effect that passage variability has on student performance. Current researchers suggest that the effects of passage variability may be mediated by administering multiple alternate passages (Poncy et al., 2005). Thus, administering multiple probes may be an additional area of future research.

The predictive utility of the MIR:R assessment system may be further established by using scores from administrations other than the sixth probe administration. Additionally, extending this research to students in grades lower and higher than grade three and to other standardized tests will be beneficial in establishing the predictive validity of the MIR:R assessment system. Finally, educators' perspectives on the practicality and utility of the MIR:R assessment system should be investigated by collecting social validity data.

In conclusion, MIR:R is a promising reading screener that can be used to predict student performance on an end-of-year, high-stakes measure for students in grade 3. MIR:R addresses some of the limitations associated with currently-available CBM measures by requiring that students simultaneously decode and comprehend connected text. MIR:R is a practical instrument that measures both reading fluency and reading comprehension in an efficient format and could be used as part of a comprehensive RTI program to identify at-risk students. Additional data

from future research will assist in determining the acceptability of the MIR:R assessment system for more widespread use by other systems.

REFERENCES

Allington, R. L. (2009). *What really matters in fluency: research-based practices across the curriculum.* Boston: Pearson Education, Inc.

Bell, S. M., Hilton-Prillhart, A. N., Hopkins, M. B., & McCallum, R. S. (2012). *Monitoring Instructional Responsiveness: Reading (MIR:R).* Unpublished test. University of Tennessee.

Bell, S. M., & McCallum, R. S. (2008). *Handbook of reading assessment.* Boston: Allyn and Bacon.

Brown, J. A., Fishco, V. V., & Hanna, G. (1993). *Nelson-Denny Reading Test: Manual for scoring and interpretation*, Forms G and H. Itasca, IL: Riverside.

Brunsman, B. & Shanahan, T. A. (2006). Review of the Dynamic Indicators of Basic Early Literacy Skills. In *Buros Mental Measurements Yearbook, 16.* Retrieved April 18, 2012, from http://web5:silverplatter.com.proxy.lib.utk.edu:90/webspirs/start.ws?customer=c14360anddatabases=S(YB).

Crawford, L., Tindal, G., & Stieber, S. (2001). Using oral reading rate to predict student performance on statewide achievement tests. *Educational Assessment, 7*, 303-323.

Daly, E. J., Chafouleas, S., & ‘’, C. H. (2005). *Interventions for reading problems: Designing and evaluating effective strategies.* New York: Guilford Press.

Deno, S. L., Fuchs, L. S., Marston, D., & Shin, J. (2001). Using curriculum-based measurement to establish growth standards for students with learning disabilities. *School Psychology Review, 30*, 507-524.

Deno, S. L., Reschly, A. L., Lembke, E. S., Magnusson, D., Callender, S. A., Windram, H., & Stachel, N. (2009). Developing a school-wide progress-monitoring system. *Psychology in the Schools, 46*, 44-55.

Dorn, S. (2010). The political dilemmas of formative assessment. *Council for Exceptional Children, 76*, 325-337.

Fuchs, L. S., & Deno, S. L. (1994). Must instructionally useful performance assessment be based in the curriculum? *Exceptional Children, 61*, 15-24.

Fuchs, L. S., Deno, S. L., & Mirkin, P. K. (1984). Effects of frequent curriculum-based measurement and evaluation of pedagogy, student achievement, and student awareness of learning. *American Educational Research Journal, 21*, 449-460.

Fuchs, L. S., Fuchs, D., Hamlett, C. L., & Stecker, P. M. (1991). Effects of curriculum-based measurement and consultation on teacher planning and student achievement in mathematics operations. *American Educational Research Journal, 28*, 617-641.

Fuchs, L. S., Fuchs, D., Hamlett, C. L., Walz, L., & Germann, G. (1993). Formative evaluation of academic progress: How much growth can we expect? *School Psychology Review, 22*, 27-48.

Good, R. H., Deno, S. L., & Fuchs, S. L. (1995). *Modeling academic growth for students with and without disabilities.* Paper presented at the third annual Pacific Coast Research Conference, Laguna Beach, CA.

Good, R. H. & Kaminski, R. A. (2001). Technical adequacy of second grade DIBELS oral reading fluency passages. *World Health Organization Technical Report Series*, *8*.

Good, R. H. & Kaminski, R. A. (Eds.) (2002). *Dynamic Indicators of Basic Early Literacy Skills* (6th ed.). Eugene, OR: Institute for the Development of Educational Achievement.

Good, R. H., & Shinn, M. R. (1990) Forecasting accuracy of slope estimates for reading curriculum-based measurement: Empirical evidence. *Behavioral Assessment, 12*, 179-194.

Hale, A. D., Skinner, C. H., Wilhoit, B., Ciancio, D., & Morrow, J. A. (in press). Variance in broad reading accounted for by measures of reading speed embedded within Maze and comprehension rate measures. *Journal of Psychoeducational Assessment*.

Hammill, D. D., Widerholt, J. L., & Allen, E. A. (2006). *Test of Silent Contextual Reading Fluency.* Austin, TX: Pro-Ed.

Harcourt Brace (2003). *Stanford Achievement Test, Tenth Edition: Technical Data Report.* San Antonio, TX: Author.

Helwig, R., & Tindal, G. (1999). *Modified measures and statewide assessments.* Unpublished manuscript.

Hiebert, E. H. (2000). *What is third-grade reading?* Paper presented at the meeting of the Southeast Literacy Consortium, Athens, GA.

Hilton-Prillhart, A. N. (2011). *Validation of the Monitoring Academic Progress: Reading (MAP: R): Development and investigation of a group-administered comprehension-based tool for RTI.* Unpublished doctoral dissertation, University of Tennessee, Knoxville.

Individuals with Disabilities Education Improvement Act of 2004, 20 U.S.C. § 1400 et seq. Stat. (2004).

Jenkins, J. R., Graff, J. J., & Miglioretti, D. L. (2009). Estimating reading growth with intermittent CBM progress monitoring. *Exceptional Children, 75*, 151-164.

Jenkins, J. & Terjeson, K. J. (2011). Monitoring reading growth: Goal setting, measurement frequency, and methods of evaluation. *Learning Disabilities Research & Practice, 26*, 28-35.

Johnson, E. S., Jenkins, J. R., Petscher, Y., & Catts, H. W. (2009). How can we improve the accuracy of screening instruments? *Learning Disabilities Research & Practice, 24*, 174-185.

Jones, E. D., & Krouse, J. P. (1988). The effectiveness of data-based instruction by student teachers in classrooms for pupils with mild handicaps. *Teacher Education and Special Education, 11*, 9-19.

Kaufman, A., & Kaufman, N. L. (2004). *Kaufman Test of Educational Achievement, Second Edition*, Form A (KTEA-II Form A). Circle Pines, MN: AGS Publishing.

Kim, D. (1993). *Trends of reading growth for students with severe reading difficulties: A four-year longitudinal study.* Unpublished doctoral dissertation, University of Minnesota, Minneapolis.

Marston, D. (1989). A curriculum-based measurement approach to assessing academic performance: What it is and why to do it. In M. Shinn (Ed.), *Curriculum-based measurement: Assessing special children* (pp. 18-78). New York: Guilford Press.

Mather, N., Hammill, D. D., Allen, E. A., & Roberts, R. (2004). *Test of Silent Word Reading Fluency.* Austin, TX: Pro-Ed.

McGill-Franzen, A., Dennis, D., Payne, R., & Solic, K. (November 2006). *Exploring the instructional utility of DIBELS as a screening and progress monitoring measure*. National Reading Conference, Los Angeles, California.

McKenna, M. C. & Stahl, S. A. (2003). *Assessment for reading instruction.* New York: Guilford Press.

Miller, N., DeLapp, R., & Driscoll, R. (2007). Group anxiety reduction in sixth grade students. *Education Resources Information Center, 11*, 1-8.

Neddenriep, C. E., Skinner, C. H., Hale, H., Oliver, R., & Winn, B. (2007). An investigation of the validity of reading comprehension rate: A direct, dynamic measure of reading comprehension. *Psychology in the Schools*, *44,* 373-388.

No Child Left Behind Act, 20 U.S. C., 2001.

Nolet, V., & McLaughlin, M. (1997). Using CBM to explore a consequential basis for the validity of a state-wide performance assessment. *Diagnostique, 22*, 146–163.

Pearson, P. D. (April 2011). *The tortured history of reading comprehension assessment*: *Are there lessons from the past? Is there hope for the future? Will we ever get it right?* American Educational Research Association, New Orleans, Louisiana.

Poncy, B. C., Skinner, C. H., & Axtell, P. K. (2005). An investigation of the reliability and standard error of measurement of words read correctly per minute using curriculum-based measurement. *Journal of Psychoeducational Assessment, 23*, 326-338.

Pressley, M., Hilden, K., & Shankland, R. (2005). An evaluation of end-grade-3 Dynamic Indicators of Basic Early Literacy Skills (DIBELS): Speed reading without comprehension, predicting little (Tech. Rep.). East Lansing, MI: Michigan State University, Literacy Achievement Research Center.

Rasinski, T. V., & Padak, N. (2004). *Effective reading strategies: Teaching children who find reading difficult* (3rd ed.). Columbus, OH: Pearson Merrill Prentice Hall.

Renaissance Learning Systems (1997). *STAR Reading Assessment*. Wisconsin Rapids, WI: Renaissance Learning Systems.

Riedel, B. W. (2007). The relation between DIBELS, reading comprehension, and vocabulary in urban first-grade students. *Reading Research Quarterly, 42*, 546-562.

Samuels, S. J. (2007). The DIBELS tests: Is speed of barking at print what we mean by reading fluency? *Reading Research Quarterly, 42*, 563–566.

Sattler, J. (2008). *Assessment of children: Cognitive foundations*. San Diego: J.M. Sattler.

Schatschneider, C., Wagner, R. K., & Crawford, E. C. (2008). The importance of measuring growth in response to intervention models: Testing a core assumption. *Learning & Individual Differences, 18*, 308-315.

Scriven, M. (1967). The methodology of evaluation. In R.E. Stake (Ed.), *Curriculum Evaluation*. Chicago: Rand McNally.

Shapiro, E. S. (2004). *Academic skills problems: Direct assessment and intervention.* The Guildford Press: New York.

Shapiro, E. S., Keller, M. A., Lutz, J. G., Santoro, L. E., & Hintze, J. M. (2006). Curriculum-based measures and performance on state assessment and standardized tests: Reading and math performance in Pennsylvania. *Journal of Psychoeducational Assessment, 24*, 19-35.

Shin, J., Deno, S. L., & Espin, C. A. (2000). Technical adequacy of the maze task for curriculum-based measurement of reading growth. *Journal of Special Education, 34*, 164-172.

Shin, J., Deno, S. L., McConnell, S. R., & Espin, C. A. (2000). *Reading-growth estimates for students in general education and students with learning disabilities using curriculum-based measurement.* Unpublished manuscript.

Shinn, M. R. (1989). *Curriculum-based measurement: Assessing special children.* New York. Guilford.

Shinn, M. R. (2002). Best practices in using curriculum-based measurement in a problem-solving model. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology, IV* (pp. 671-697). Bethesda, MD: National Association of School Psychologists.

Shinn, M. R., Good, R. H., Knutson, N., & Tilly, W. D. (1992). Curriculum-based measurement of oral reading fluency: A confirmatory analysis of its relation to reading. *School Psychology Review, 21*, 459-479.

Shinn, M. M. & Shinn, M. R. (2002). *AIMSweb training workbook.* San Antonio: NCS Pearson.

Shinn, M. R., Shinn, M. M., Hamilton, C., & Clarke, B. (2002). Using curriculum-based measurement in general education classrooms to promote reading success. In M.R. Shinn, H.M. Walker, & G. Stoner (Eds.), *Interventions for academic and behavior problems II: Preventive and remedial approaches* (pp. 113-142). Bethesda, MD: National Association of School Psychologists.

Silberglitt, B., & Hintze, J. M. (2005). Formative assessment using CBM-R cut scores to track progress toward success on state-mandated achievement tests: A comparison of methods. *Journal of Psychoeducational Assessment, 23*, 304-325.

Silberglitt, B., & Hintze, J. M. (2007). How much growth can we expect? A conditional analysis of R-CBM growth rates by level of performance. *Council for Exceptional Children, 74*, 71-84.

Skinner, C. H., Neddenriep, C. E., Bradley-Klug, K. L., & Ziemann, J. M. (2002). Advances in curriculum-based measurement: Alternative rate measures for assessing reading skills in pre- and advanced readers. *Behavior Analyst Today, 3*, 270-281.

Skinner, C. H., Williams, J. L., Morrow, J. A., Hale, A. D., Neddenriep, C., & Hawkins, R. O. (2009). The validity of a reading comprehension rate: Reading speed, comprehension, and comprehension rates. *Psychology in the Schools*, *46,* 1036-1047.

Stage, S. A. & Jacobsen, M. D. (2001). Predicting student success on a state-mandated performance-based assessment using oral reading fluency. *School Psychology Review, 30*, 407-419.

Tennessee State Department of Education. *Tennessee Comprehensive Assessment Program*. Published test. State of Tennessee. Retrieved February 1, 2012, from http://tn.gov/education/assessment/achievement.shtml.

Thurlow, M. L., & Thompson, S. J. (1999). District and state standards and assessments: Building an inclusive accountability system. *Journal of Special Education Leadership, 12*, 3-10.

Torgensen, J. K., Wagner, R. K., & Rashotte, C. A. (1999). *TOWRE: Test of Word Reading Efficiency: Examiner's manual.* Austin, TX: Pro-Ed, Inc.

Wayman, M. M., Wallace, T., Wiley, H. I., Ticha, R., & Espin, C. A. (2007). Literature synthesis on curriculum-based measurement in reading. *The Journal of Special Education, 41*, 85-120.

Wiederholt, J. L., & Bryant, B. R. (2001). *Gray Oral Reading Tests – Diagnostic.* Los Angeles, CA: Western Psychological Services.

Wiliam, D. (2006). Formative assessment: Getting the focus right. *Educational Assessment, 11*, 283-289.

Williams, J. L., Skinner, C. H., Floyd, R. G., Hale, A. D., & Neddenriep, C., & Kirk, E. (2011). Words correct per minute: The variance in standardized reading scores accounted for by reading speed. *Psychology in the Schools*, *48,* 87-101.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock-Johnson III Tests of Achievement-Revised.* Itasca, IL: Riverside.

Yeo, S., Fearrington, J. Y., & Christ, T. J. (2012). Relation between CBM-R and CBM-mR slopes: An application of latent growth modeling. *Assessment for Effective Intervention, 37*, 147-158.

Zhou, X. H, Obuchowski, N. A., & Obushcowski, D. M. (2002). *Statistical methods in diagnostic medicine.* New York: Wiley.

APPENDIX A

Table 1

Alternate Form Reliability – Monitoring Instructional Responsiveness: Reading (MIR:R)

	Probe 1	*Probe 2*	*Probe 3*	*Probe 4*	*Probe 5*	*Probe 6*	*Probe 7*	*Probe 8*	*Probe 9*
US1	.77	.68	.61	.66	.71	.72	.70	.72	.70
Probe 1		.76	.70	.72	.75	.76	.74	.77	.75
Probe 2			.77	.76	.73	.75	.73	.73	.73
Probe 3				.78	.75	.74	.72	.73	.74
Probe 4					.80	.77	.77	.76	.75
Probe 5						.81	.77	.80	.80
Probe 6							.81	.80	.81
Probe 7								.84	.84
Probe 8									.85
Probe 9									

Note. All correlations are significant at $p < .01$.

Table 2

Descriptive Statistics and t Tests for Monitoring Instructional Responsiveness: Reading (MIR:R) Slopes

		n	*M*	*SD*	*M* Difference	*SD* Difference	*t*	*df*	*p*
	Slope 1 -	477	3.51	6.10					
Pair 1	Slope 2	477	4.33	7.63	-.82	4.55	-3.94	476	< .001
	Slope 1 -	470	3.59	6.07					
Pair 2	Slope 3	470	2.79	7.47	.80	4.81	3.61	469	< .001
	Slope 1 -	473	3.53	6.12					
Pair 3	Slope 4	473	4.18	7.01	-.64	3.92	-3.55	472	< .001

Table 3

Descriptive Statistics for Monitoring Instructional Responsiveness: Reading (MIR:R), Sixth Probe Administration, and Tennessee Comprehensive Assessment Program (TCAP), Achievement Test, Reading Composite Scores

	n	*M*	*SD*
Comprehension Percentage	457	50.73	30.70
Total Words Read	457	250.61	86.01
Comprehension Rate Static Score	457	124.91	86.31
Comprehension Rate Slope Score	478	3.51	6.10
TCAP Scale Score	469	749.17	33.20

Table 4

Correlation Coefficients Among Monitoring Instructional Responsiveness: Reading (MIR:R) Comprehension Percentage, Total Words Read, Comprehension Rate Static, Comprehension Rate Slope, and Tennessee Comprehensive Assessment Program (TCAP), Achievement Test, Reading Composite Scores

	Comprehension Percentage	Total Words Read	Comprehension Rate Static	Comprehension Rate Slope	TCAP Scale Score
Comprehension Percentage		-.16*	.54*	.19*	.31*
Total Words Read			.27*	.07	.06
Comprehension Rate Static				.35*	.60*
Comprehension Rate Slope					.22*
TCAP Scale Score					

Note. *Correlation is significant at $p = .01$.

Table 5

Sensitivity and Specificity Information for Monitoring Instructional Responsiveness: Reading (MIR:R) Comprehension Rate Static Score to Predict Tennessee Comprehensive Assessment Program (TCAP), Achievement Test, Reading Composite Non-Proficiency Status

	Sensitivity	*Specificity*	*Cut Score**	*ROC AUC*	*TP*	*FP*	*TN*	*FN*	*Classification Accuracy*
Comprehension Rate Static Score	90	55	175.55	.81	227	69	110	43	75%

Note. *Cut scores were identified that resulted in sensitivity levels as close to 90% as possible.

Table 6

Sensitivity and Specificity Information for Monitoring Instructional Responsiveness: Reading (MIR:R) Comprehension Rate Slope to Predict Tennessee Comprehensive Assessment Program (TCAP), Achievement Test, Reading Composite Non-Proficiency Status

	Sensitivity	*Specificity*	*Cut Score**	*ROC AUC*	*TP*	*FP*	*TN*	*FN*	*Classification Accuracy*
Comprehension Rate Slope	90	24	9.65	.62	254	140	45	28	64%

Note. *Cut scores were identified that resulted in sensitivity levels as close to 90% as possible.

Table 7

Step-wise Multiple Regression Analysis Predicting Tennessee Comprehensive Assessment Program (TCAP), Achievement Test, Reading Composite, with Monitoring Instructional Responsiveness: Reading (MIR:R) Comprehension Percentage and Total Words Read Component Scores

Predictor	R^2 *Change*	*SEb*	β	*F Change*	*p*
Comprehension Percentage	.094	2.68	.31	46.55	< .001
Total Words Read	.011	.02	.11	5.66	= .018

Table 8

Sensitivity and Specificity Information for Monitoring Instructional Responsiveness: Reading (MIR:R) Comprehension Percentage and Total Words Read Component Scores to Predict Tennessee Comprehensive Assessment Program (TCAP), Achievement Test, Reading Composite Non-Proficiency Status

	Sensitivity	*Specificity*	*Cut Score**	*ROC AUC*	*TP*	*FP*	*TN*	*FN*	*Classification Accuracy*
Comprehension Percentage	90	52	74.54	.80	228	70	109	43	75%
Total Words Read	90	.06	396.50	.55	238	72	107	33	77%

Note. *Cut scores were identified that resulted in sensitivity levels as close to 90% as possible.

Table 9

Step-wise Multiple Regression Analysis Predicting Tennessee Comprehensive Assessment Program (TCAP), Achievement Test, Reading Composite, with Monitoring Instructional Responsiveness: Reading (MIR:R) Comprehension Percentage, Total Words Read, and Comprehension Rate Slope

Predictor	R^2 *Change*	*SEb*	β	*F Change*	*p*
Comprehension Percentage	.093	2.69	.31	46.04	< .001
Comprehension Rate Slope	.024	.26	.16	12.02	= .001
Total Words Read	.008	.02	.09	4.14	= .042

VITA

Kelli Caldwell Miller graduated from the University of South Carolina, Columbia, in May 2008 with a Bachelor of Arts degree in Experimental Psychology, with a minor in General Education. Subsequently, Kelli attended the University of Tennessee, Knoxville, and earned a Master's of Science degree in Applied Educational Psychology in December 2011. Kelli is currently completing a pre-doctoral internship at the Tennessee Internship Consortium. She will graduate from the University of Tennessee, Knoxville, with a Doctor of Philosophy degree in School Psychology in August 2013.